ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

FISICA

Ciclo 37

Settore Concorsuale: 02/D1 - FISICA APPLICATA, DIDATTICA E STORIA DELLA FISICA

Settore Scientifico Disciplinare: FIS/07 - FISICA APPLICATA A BENI CULTURALI, AMBIENTALI, BIOLOGIA E MEDICINA

Machine Learning and network-based methods for the analysis of biological systems involved in food processing

Presentata da: Anis Mansouri

Coordinatore Dottorato

Prof. Alessandro Gabrielli

Supervisore

Prof. Daniel Remondini

Esame finale anno 2025

Contents

The E-MUSE Project

"European dairy industry is an important agri-food sector; it represents more than 300,000 jobs and 10 billion € positive trade balance. Five out of ten top global dairy companies are European and more than 80 % of them are SMEs. More than 300 European cheeses and dairy products are sold all over the world and are protected as geographical indications or traditional specialties. Mastering a cheese-ripening process to avoid sanitary risk and waste, and produce typical cheeses with organoleptic properties valued by the consumers is of economic and social significance.

Complex microbial Ecosystems MUltiScale modElling (E-MUSE) project aims to develop innovative modelling methodologies to improve knowledge about complex biological systems and to control and/or predict their evolution by combining artificial intelligence and systems biology. This multidisciplinary strategy integrating genome-scale metabolic models, dynamic modelling methodologies, together with the design of efficient statistical and machine learning tools, will allow analysis of multi-omics data and application of the results to macro-scale properties related to cheese ripening and consumer preference.

Moreover, in the context of sustainable development, more and more consumers are diversifying their diet and consume plant-based food. Introduction of plant-based proteins in the cheese process brings issues such as acidity or safety. Modelling strategies of E-MUSE will help to target and solve these issues. E-MUSE will train researchers with multidisciplinary skills in mathematics, bioinformatics and biology to design and use innovative multiscale modelling methodologies, giving researchers a harmonised language to address future research questions about complex biological systems.

Finally, the ultimate outcome of E-MUSE is to develop, for the industry, a dynamic modelling software to control the food process." (source: https://www.itn-emuse.com)

Our collaboration within the E-MUSE Project

We owe the mouth-watering Parmigiano that we enjoy with a glass of good wine to tiny microscopic artisans, which cooperate harmoniously to craft delicious and healthy food products.

Microorganisms have specific but complementary and interdependent roles during food-making processes. They create functional networks where different species/strains share many substrates, products, and enzymes, and have many tasks in common. On the other hand and at the sub-cellular level, each specific microorganism has its own molecular networks that ensure its functions, like signaling networks, regulatory networks and metabolic networks involving enzymes and metabolites.

In order to study these inter and intra-species interactions and predict their outcome, we used machine learning and network theory approaches to reconstruct and analyze these networks. These approaches allow the integration of multi-omics data (genomics, transcriptomics, metabolomics...) and, thus, allow a more realistic modeling and representation of the complex biological processes that govern food making.

Acknowledgment

This thesis and its objectives wouldn't have been achieved without the valuable supervision of Prof. Daniel Remondini and co-supervision of Prof. Enrico Giampieri from the University of Bologna. Throughout my thesis, Prof. Remondini provided me with

relevant instructions as well as brainstorming ideas which were the basis of all the work and the results we achieved.

I have also to thank Prof. Eva Balsa-Canto, Prof. Bas Teusink, Prof Vidács László and Prof. Péter Pusztai for hosting me within their groups where I learned valuable knowledge and skills in computational systems biology and machine learning. Finally, I thank all the E-MUSE consortium headed by Dr. Dominique Swennen for their significant contribution to my personal and professional development.

Funding

The entire thesis project was funded by E-MUSE MSCA-ITN-2020 European Training Network under the Marie Sklodowska-Curie grant agreement No. 956126. The funding did not influence the study's design, the interpretation of data, or the writing of the manuscript.

Key words and symbols

SLAB:	Starter Lactic Acid Bacteria
NSLAB:	Non-Starter Lactic Acid Bacteria
E. coli:	*Escherichia coli*
AMR:	Antimicrobial Resistance
RF:	Random Forest
EN:	Elastic Net
ND:	Network Diffusion
PPI:	Protein–Protein Interaction network
CD:	Community Detection
IVI:	Integrative Value of Influence
CM:	Centrality Measure
DC:	Degree Centrality
BC:	Betweenness Centrality
CC:	ClusterRank Centrality
LHC:	Local H Index
NC:	Neighborhood Connectivity
CIC:	Collective Influence Centrality
GSMM:	Genome-Scale Metabolic Model
FBA:	Flux Balance Analysis
FVA:	Flux Variability Analysis

Chapter 1

Introduction

In this introductory chapter, we introduce the different methods and topics addressed during this thesis, i.e:

- The use of microbial gene expression as a proxy to estimate flavor levels in cheese (Chapter 2).
- The use of network approaches to identify new genes associated to antimicrobial resistance in *Escherichia coli* (Chapter 3).
- The combination of network-based and flux-based analyses to create context-specific metabolic networks of yeasts (Chapter 4). This work is in collaboration with PhD. Diego Troitiño from the Bio2Eng group headed by Prof. Eva Balsa-Canto.

The methods used and developed throughout this thesis can be applicable in different contexts. Here, we apply them in our main research topics which are related to food sector, namely, food making and food safety. In the introduction section of chapters 2, 3 and 4, we report some other research fields where similar methods have been already used.

1.1 Cheese: an overview

In this section, we introduce the biochemistry and the microbiology of cheese. We present basic biological and biochemical facts that help the non-specialist reader to understand the motivations and the results of our work described in Chapter 2. In this work, we analyzed multi-omics data including metatransciptomics, metametabolomics and growth data collected from an experimental surface-ripened cheese. We first used metatranscriptomics and metametabolomics to train classification models that infer metabolites levels from microbial gene expression. The features selected and used to construct the models and perform the predictions were biologically relevant as they were consistent with biological metabolic pathways. We also investigated the contribution of each microbial species to these features, i.e, which feature(s) (gene(s)) belongs to which microbe, and we found that bacteria contributed more than yeasts. To check this finding, we performed correlation analyses to study the association between microbial growth and the metabolites and, indeed, we found strong correlations between the bacteria and most of the metabolites, contrary to yeasts which had poor correlations. Yeasts play an important role in flavor development in real cheeses, but as the studied cheese in this work is experimental, we couldn't detect a significant yeast-metabolite correlation. Nevertheless, the correlation results are still in concordance with the results of the modeling step.

1.1.1 Brief history

About nine thousands years ago, human began the domestication of milk-producing animals. Archaeologist argued that these domestic animals were initially kept only for their meat, bones and hide, and that additional uses of these animals have began later.

Around 3500 BC, the practice of keeping mammals for their secondary products and uses such as milk, wool and labor has been adopted and spread through western Asia, Europe and as far east as India [1]. Milk was consumed fresh due to its instability, especially in hot conditions, where it spoils and sours quickly. The curdling of milk occurs naturally by the effect of microbes such as bacteria and yeasts. Humans took advantage of this natural phenomenon to preserve milk, i.e, transform it into cheese, a more stable product which preserves the nutritional benefits of milk on top of its appealing gustatory characteristics. The earliest surviving cheese was found by archaeologists in a tomb from the Egyptian First Dynasty between 3100 and 2900 BC. In the same epoch, the word "cheese" was reported in Sumerian civilization's literature, mainly, in food glossaries including a list of twenty distinct cheeses [2]. Later, cheese started expanding westward. It was made throughout western and central Europe countries such as Greece, Austria and Switzerland. During the Roman Empire, cheese ("caseus") was a favorite luxury food, thus, its production and trading have grown in many Mediterranean countries from Spain, France to Turkey. It's worth-noting that cheese was also known in Eurasian and central Asia countries such as Russia and Mongolia [3].

Nowadays, there are at least 1000 varieties of cheese [4], around one third are already described, classified and named [5, 6]. The Food and Agriculture Organization of the United Nations reported that cheese exports reached 2.6 million tonnes in 2019 [7]. In 2020, EU countries were the largest cheese producers and consumers, followed by the United States and Canada [8].

1.1.2 Biochemistry

Cheese results from a transformation of raw milk by different microbial species such as bacteria, yeasts and molds. Cheese production relies on six major steps: (1) milk selection, (2) acidification, (3) coagulation, (4) dehydration of the coagulum to obtain the curd, (5) shaping of the curd, (6) ripening (maturation) of the curd to obtain a desired texture and flavor profile [9]. At the molecular level, raw milk components such as lactose, lipids and proteins undergo primary and secondary events [10]. Primary events involve:

- **Metabolism of lactose** present in the raw milk by lactic acid bacteria (starters). The conversion of lactose into lactate causes the acidification of the milk which prevents spoilage by pathogenic organisms. The starters also produce enzymes responsible for the metabolism of residual lactate and citrate, which are very important substrate for a range of reactions during ripening, in addition of being precursors for flavor compounds in some cheese varieties.

- **Lipolysis** of milk triglycerides by indigenous, endogenous and/or exogenous lipase to fatty acids which are important precursors for the production of volatile flavor compounds.

- **Proteolysis** is the most complex and important of the primary biochemical events. It occurs by proteolytic action of endogenous enzymes as well as proteinases and peptidases from the starters, non-starter lactic acid bacteria and perhaps secondary microflora, to produce free amino acids which are important precursors of many flavor compounds.

The molecules resulting from the primary events undergo **secondary biochemical events** consisting mainly in the metabolism of free fatty and amino acids leading to the formation of flavor compounds. Free fatty acids can directly contribute to cheese flavor. Free amino acids as well as fatty acids can act as precursors for reactions leading to the formation of flavors.

1.1.3 Microbiology

Microbes involved in cheese making can be classified into two categories: Starter Lactic Acid Bacteria (SLAB) and secondary microbiota. Both SLAB and secondary microbiota can be intentionally introduced into the cheese or accidentally through a contact

with the cheese making equipment. SLAB, such as *Lactococcus and Streptococcus thermophilus*, are mainly used to start the ripening process by acidifying the milk, but can also contribute to flavor compound formation during cheese ripening. The secondary microbiota includes four groups: (1) Non-Starter Lactic Acid Bacteria (NSLAB) such as nonstarter *lactobacilli, Pediococcus, Enterococcus,* and *Leuconostoc*; (2) propionic acid bacteria; (3) molds; and (4) bacteria and yeast, which grow on the surface of smear-ripened cheeses. The secondary microbiota doesn't play a significant role in cheese acidification but has an important role in flavor development during cheese ripening [11].

1.1.4 Multi-omics data integration for the characterization of cheese

The transformation of milk into cheese requires a complex set of metabolic reactions. Each reaction is catalyzed by one or several enzymes to produce certain products from the substrates available in milk. These enzymes can be endogenous, i.e, already present in the milk such as proteases and lypases, or expressed by the microbes involved in cheese fermentation and ripening, such as the SLAB and the secondary microflora. The different molecules resulting from this complex dynamics and biochemical reactions can be measured thanks to omics approaches that are able to collect multiple layers of information such as (meta)genomics, (meta)transcriptomics, (meta)proteomics and (meta)metabolomics. Thus, an integrative analysis of the different omics-data would reveal more biological insights about the mechanisms that govern cheese ripening.

Microbes play a crucial role in cheese ripening and there are tight inter-relationships between the characteristics of cheese such as texture and flavor and its microbial composition. Microbiome-Metabolome inter-relationships are of great interest in cheese industry as flavors and aroma are the main properties that influence consumers' preferences [12, 13, 14]. Many studies have been conducted to analyze these relationships and try to define fingerprints allowing the characterization of cheese. Afshari et al. [15, 16] analyzed metagenomics and metabolomics data collected from artisanal and industrial Cheddar cheese of different brands and age. They could identify dominant taxa and metabolites that can differentiate artisanal Cheddar from the industrial one. In addition, they could find a correlation between OTUs[1], the metabolites and the cheese type as some OTUs and metabolites were strongly associated to a given cheese type, its brand and its ripening age. These studies identified biomarkers (OTUs, metabolites...) that can be used to determine cheese quality and authenticity as they allow to distinguish between different cheese types and brands. Bertuzzi et al. [17] studied the association between flavor development and the microbial composition in surface-ripened cheese. They could find correlations between bacteria and yeasts' abundance and the levels of the major volatile compounds classes such as alcohols, aldehydes, esters...

In Chapter 2, we present a machine learning approach to investigate the microbiome-metabolome relationship in an artificial surface-ripened cheese.

1.2 Network science

In this section, we present the basic concepts in network science, i.e, a definition of a graph (network), the types of graphs, their representation and the different metrics and methods commonly used to extract information from them (centrality measures, node prioritization, community detection...). We applied two network-based methods: 1) Network diffusion to prioritize genes eventually associated to antimicrobial resistance in *Escherichia coli* (Chapter 3); 2) Centrality measures to detect relevant reactions used to build reduced and context-specific yeast genome scale metabolic model (Chapter 4).

[1]Operational Taxonomic Unit (OTU): a group of closely related microbes which are arranged together based on the similarity of specific genomic sequences.

1.2.1 Foundation

Graph theory or network theory has been invented by the famous mathematician Leonhard Euler in the 18^{th} century. The story of network science started in Kaliningrad city (previously called Königsberg) in Russia. In this city, there were four districts connected to each other by seven bridges, every Sunday, the citizens of Kaliningrad enjoyed playing a challenging game: it was a kind of puzzle consisting in walking through all the part of the city in a continuous walk while crossing each of the seven bridges only once, and return to the starting point. No one could find such a path and the problem was then called "The Seven Bridges of Königsberg". When this came to Eular attention, he applied himself to solve it. In 1741, he published an article [18] where he provided mathematical demonstration to prove that the problem is unsolvable, and provided also a general rule to answer the question whatever the number of bridges. Euler's work has been considered as the earliest in graph science [19, 20, 21].

1.2.2 Basic concepts

Definition

A graph $\mathbf{G}$ is a representation of associations between a finite set of objects. It's a diagram where each object called "node" or "vertex" $\mathbf{V}$ is connected to other nodes by one or more links called "edges" $\mathbf{E}$:

$$\mathbf{G} = \mathbf{V}.\mathbf{E}$$

$\mathbf{V}$ is a vector of vertices $\mathbf{V} = \{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_n}\}$ and $\mathbf{E} = \{\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_n}\}$ is a vector of edges. An edge $\mathbf{e_k} = (\mathbf{v_i}, \mathbf{v_j})$ connects vertices $\mathbf{v_i}$ and $\mathbf{v_j}$.

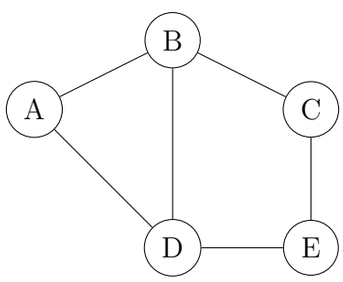


Figure 1.1: Undirected simple graph with 5 nodes and 6 edges.

In the graph (Figure 1.1), $\mathbf{V} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}\}$ and $\mathbf{E} = \{(\mathbf{A}, \mathbf{B}), (\mathbf{A}, \mathbf{D}), (\mathbf{B}, \mathbf{A}), (\mathbf{B}, \mathbf{C})$ $(\mathbf{B}, \mathbf{D}), (\mathbf{C}, \mathbf{B}), (\mathbf{C}, \mathbf{E}), (\mathbf{D}, \mathbf{E})\}$. Graph $\mathbf{G}$ is an undirected graph because the nodes are not associated with directions, i.e, the edges $(\mathbf{v_i}, \mathbf{v_j}) = (\mathbf{v_j}, \mathbf{v_i})$.

Based on the characteristics of their edges, graphs can be classified as follows (see Figure 1.2):

- **Simple graphs:** node pairs are connected by only one edge and no node connects with itself through a loop edge.
- **Multiple graphs:** node pairs can be connected by more than one edge (multiple edges) and loop edges can exist.
- **Directed graphs:** edges have heads and tails, i.e, they are directed from one node $\mathbf{v_i}$ (the tail) to the other node $\mathbf{v_j}$ (head). Unlike undirected graphs, for any edge $\mathbf{v_i} \rightarrow \mathbf{v_j}$, the edges $(\mathbf{v_i}, \mathbf{v_j}) \neq (\mathbf{v_j}, \mathbf{v_i})$.
- **Weighted graphs:** edges are weighted by a numerical value that represents the strength of the association between node pairs, but, depending on the context, it can have different meanings, e.g, in a social network, the weight can represent the number of common friends between two individuals, in a gene co-expression network, the edges' weights represent the correlation between the genes.

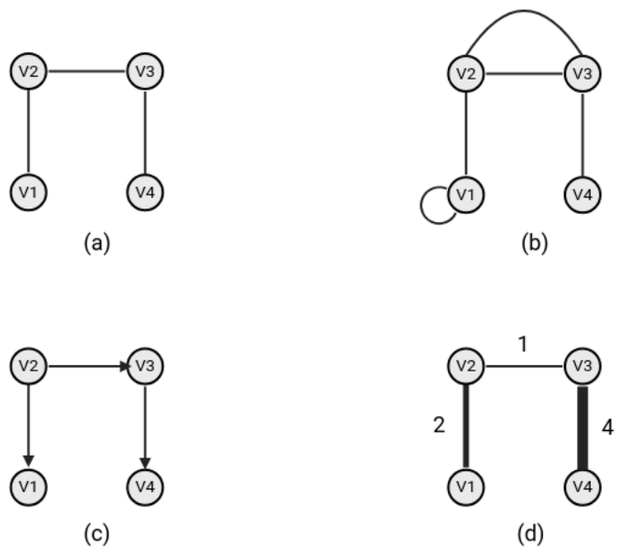


Figure 1.2: (a) Simple graph; (b) Multiple graph; (c) Directed graph; (d) Weighted graph.

Graphs can be also classified according to their node types:

- **Unipartite graphs:** also called one-mode graphs. The nodes in these graphs are of the same type, e.g, a gene network is unipartite because all nodes are genes.
- **Bipartite graphs:** also called bigraphs, they contain two types of nodes and a set of edges connecting only nodes of different types [22], e.g, metabolic networks are bipartite because nodes can be either reactions or metabolites, and edges represent the association between reactions and metabolites.

Representation of graphs

Graphs are represented and maintained in the computer memory in the form of two standard data structures: arrays and lists [23, 24]. Array representation of a graph consists in a two-dimensional square matrix called adjacency matrix (Table 1.1). For a simple undirected and unweighted graph having $\mathbf{m}$ nodes, the adjacency matrix $\mathbf{A} = [\mathbf{a_{ij}}]$ is a symmetric $\mathbf{m} \times \mathbf{m}$ matrix. At each position $(\mathbf{v_i}, \mathbf{v_j})$, the entry can be either 1 or 0 representing connection (1) or absence of connection (0) between $\mathbf{v_i}$ and $\mathbf{v_j}$:

$$a_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Secondly, graph can be represented by a two-column edge list (Table 1.2). Conventionally, edges go from the node in the first column to the node in the second one in the case of directed networks. If the network is weighted, a third column containing the edge's weight is added.

	A	B	C	D	E
A	0	1	0	1	0
B	1	0	1	1	0
C	0	1	0	0	1
D	1	1	0	0	1
E	0	0	1	1	0

Table 1.1: Adjacency matrix of the graph in figure 1.1

A	B
A	D
B	C
B	D
C	E
D	E

Table 1.2: Edge list of the graph in figure 1.1

For relatively small graphs, adjacency matrices are easily accessed and traversed. However, in the case of large graphs or sparse graphs containing more nodes than edges, the use of array representation is ineffective due to high memory requirement. In such cases, list representation is a good memory-effective alternative.

1.2.3 Network centrality measures

Many social, economic or biological processes can be modeled as networks of connected nodes (individuals, companies, biomolecules...). The need of metrics that can measure the importance or the influence of the actors (nodes) within those networks raised a rush of research as they can provide an effective tool to understand networks and extract useful information from them. In the late 1940s, the Group Networks Laboratory at MIT directed by Alex Bavelas has made the first attempts to define centrality indices to measure the individuals' implication in information flow within professional and social networks [25]. Afterwards, various centrality measures (CMs) have been developed and used depending on the context and the applications. For a comprehensive review of CMs, please refer to these research articles [26, 27, 28, 29, 30]. Here, we describe eight of the most commonly used CMs:

- **Degree:** also called *connectivity*, it is the simplest local CM. It reflects how well a node is connected within the network, i.e, how many links it has with other nodes [31, 32]. In weighted networks, the degree of a node is a float called *strength*. In directed networks, In-Degree and Out-Degree represent the number of edges directed into and out of a node, respectively.

- **Betweenness:** introduced by Freeman 1977 [33], it measures how many times a node lies on the shortest path[2] between node pairs. Nodes having a high betweenness play a role of "bridges", i.e, it's often needed to pass through them to go from a node to another, or from a group of nodes to another while covering the shortest distance. For example, in a network of individuals, people having a high betweenness play an important role in transferring information between groups.

- **Clustering coefficient:** measures the tendency of the nodes in the network to cluster together. There exist two versions: Local and Global clustering coefficient (CF). Local CF is calculated by dividing the number of a node's links within its neighborhood by the number of links that could possibly exist between them [34]. Global CF (a.k.a. transitivity) indicates the clustering of the whole network. It's the proportion of closed triplets [3] out of all triplets in the network [35].

- **ClusterRank:** accounts for the node's influence (the degree) as well as its neighbors' influences in addition to it's local clustering coefficient [36].

- **H index:** is a score proposed by Jorge Hirsch to assess the productivity and the impact of scientists or scholars [37, 38]. It was then introduced into network science as a centrality measure by Korn et al [39]. For example, a node having H-index = 4 means that it's connected to 4 nodes, each one of them is linked to, at least, 4 other nodes. H-index has been used to measure the nodes' influence in many real-world networks [40, 41].

- **Local H index:** H-Index has a resolution limit as it assigns the same score to too many nodes. To overcome this problem, an improved version called Local H-Index was designed [42] to take into account both H-index of the node and the H-Index of its neighbors. That is to say, the higher the H-Index of a node's neighbors, the more influential is the node.

- **Neighborhood Connectivity:** is defined as the average connectivity of a node's neighbors [43]. The importance of a node doesn't rely only on it's degree (connectivity) but also on the connectivity of its neighbors.

- **Collective influence:** is calculated by multiplying a node's reduced degree (degree - 1) by the reduced degrees of all nodes at a distance l from it, with l being the shortest path around that node. CL is a global centrality metric which

[2]Shortest path: the path with the minimum number of edges between two nodes.
[3]closed triplets: three nodes all connected together by three edges.

relies on the effect of node removal on the entire network (percolation): (1) the CLs of all nodes are calculated, then, nodes with the highest CL are removed; (2) CL recalculated again to find the new high-scored nodes to remove; (3) the process is reiterated until the giant component of the network vanishes. Nodes that causes the dismantling of the giant component are considered as influential nodes [44].

CMs are computed directly from adjacency matrices or edge lists and can be categorized into three categories:

- **Global CMs:** called "global" because the entire (global) structure of the network is needed to compute them, e.g, Closeness CM [25, 45] which reflects how close/far is a node from the other nodes in a network. A node having high closeness has an important role in spreading information within the network as it can quickly reach many nodes. In the case of large networks, computing global CMs has high computational complexity.

- **Local CMs:** don't require the entire network to be calculated. They take into consideration only the node's first neighborhood such as Degree CM.

- **Semi-Local CMs:** intermediate between global and local CMs as they encompass an environment that is farther than the node's first neighbors, e.g, Neighborhood Connectivity CM.

Centrality measures are used in a multitude of contexts. Wang et al. [46] studied nodal centrality of China's air transport network and they found a correlation between the socio-economic indicators of cities and their centralities. In biological contexts, Özgür et al. [47] could identify genes related to prostate cancer by means of a centrality-based ranking performed on gene networks. Martino et al. [48] used degree and eigenvector centrality to study attention-deficit/hyperactivity disorder on the brain connectivity network.

Each CM reflects a specific characteristic of the node. To extract as much information as possible about the nodes in a network, CMs shouldn't be considered individually, but combined to synergize their effects. In Chapter 4, we discuss the methods that can be used to combine CMs.

1.2.4 Network diffusion

Network diffusion (ND) also called network propagation relies on the propagation of an information from source node (seed node) to other nodes in a given network under the assumption that nodes underlying similar characteristics tend to interact with one another [49]. In an iterative manner, the information is propagated through the network, and, after convergence, the non-source nodes get a propagation (diffusion) score reflecting how likely they share the characteristics of the source nodes. For instance, ND was used to prioritize genes related to hereditary disorders in protein–protein interaction networks (PPI). The phenotype (information) was propagated from genes already known to be related to hereditary disorders (the seed genes) to the other uncharacterized genes within the network [50] in order to obtain a list of new genes potentially associated to the disease. Many diffusion algorithms have been developed and used in different contexts. In Chapter 3, we used the Bersanelli et al. method [51] to prioritize genes potentially associated to antimicrobial resistance in *E. coli*.

1.2.5 Community detection in networks

Community detection (CD) also called module or cluster detection is another network-based method applied in different contexts such as sociology, economy, technology and biology. Communities are clusters of nodes tending to interact more with each other as compared to the other nodes in the network. Those nodes probably have similar properties, thus, the formed cluster may have a specific function(s) within the network.

Therefore, CD is of great importance as it can reveal additional information about the network which cannot be uncovered if the nodes were analyzed individually [52, 53].

The first analyses of communities have been performed on social networks. In 1927, Stuart Rice [54] has designed an index of cohesion and an index of likeness for a quantitative measure of the voting pattern within and between groups of people, respectively. He applied his measures to small political bodies and could cluster people according to their voting behavior. Later, Weiss and Jacobson [55] introduced a CD method which is considered as an early version of several modern CD algorithms. The authors studied working relationships between 196 members (nodes) of a governmental agency (the network). By means of private interviews, the authors could report 2400 work relationships (edges) between the members of the agency. Finally, to define the workers clusters, they omitted the members having work relationships with different groups. Those members played a role of bridges and their removal permitted the separation of the distinct work groups.

Later on, many algorithms have been developed to perform CD and, more importantly, metrics have been designed to assess the clustering quality to help the algorithms find the optimal clusters. The quality function mostly used in many algorithms is *Modularity*. *Modularity* a is quantitative measure (a score) introduced by Newman and Girvan [56] to quantify how tightly the nodes are connected to each other within a defined module (cluster). It relies on the assumption that random graphs are not expected to have a cluster structure, therefore, one can detect communities on a graph by comparing its edges' density to the density of the same edges in a random graph (null model). The function is formulated as follows:

$$Modularity = \frac{1}{2m} \sum_{ij} (A_{ij} - R_{ij})\, \delta(C_i, C_j)$$

where

$$R_{ij} = \frac{k_i k_j}{2m}$$

and m is the total number of edges in the graph; A is the adjacency matrix of the graph we want to cluster. k_i and k_j are degrees of nodes i and j, respectively. R_{ij} computes the number of edges between i and j in the random graph (the null model). δ is a function that renders 1 if the nodes i and j are in the same community ($C_i = C_j$), 0 otherwise.

Modularity is used as a quality function, i.e, it measures how good/bad are the results returned by CD algorithms. It is an index to assess the quality and the relevance of the clustering. For an extensive description of various CD methods and quality functions, please refer to [57, 58, 59, 60]. Here, we present three categories of CD algorithms used in social and biological contexts and their basic principles:

- **Similarity-based methods:** create clusters based on the (diss)similarity between the nodes, such as graph partitioning [61, 62] and hierarchical clustering [63].

- **Modularity-based algorithms:** rely on the optimization of modularity to define the clusters, such as Newman's greedy optimization algorithm [64] and Louvain algorithm [65]. These algorithms combine hierarchical clustering and modularity such that nodes are grouped together in a way that maximizes their modularity (more detailed description in Chapter 3).

- **Divisive algorithms:** rely on the division of the network into communities by removing the edges that connect many clusters. Girvan and Newman [66, 56] developed a popular algorithm to find these inter-community edges. To define the underlying community structures, the algorithm removes the edge having a high betweenness, i.e, a high number of shortest paths that go through them.

1.3 Antimicrobial resistance

Bacterial infections especially those caused by contaminated food represented for us an interesting topic to address in this thesis. We were interested in the Gram-negative bacteria *Escherichia coli* as it's associated to many infectious disease in humans [67]. Different antibiotic resistant *E coli* strains are already identified in contaminated food such as meat, eggs and milk [68], therefore, we wanted to investigate more *E coli's* resistance to antibiotics for a better characterization that could contribute to the identification of potential drug targets.

In Chapter 3, we describe a systems biology approach based on network and statistical methods to analyses the *E coli's* PPI. By means of network diffusion, we could prioritize a list of genes potentially associated to AMR. Then, we could experimentally validate four candidate genes that showed significant resistance against six antimicrobials.

1.3.1 Overview

Bacteria and archaea are the earliest living organisms on Earth which appeared around 3.6 billions years ago [69]. Despite their low complexity, bacteria developed an effective strategy to survive under stressful conditions. Similar to an immune system in animals, bacteria can protect themselves by resisting and bypassing the effect of harmful agents such as antimicrobials. Antimicrobial resistance (AMR) was firstly identified in 1940 in *Staphylococcus* against penicillin-R [70], however, later scientific expeditions have demonstrated that AMR predated the introduction of antimicrobials by humans. In Ellesmere Island in Canada, scientists have discovered 2000-years-old microbial isolates that carry AMR against ampicilin [71]. Furthermore, targeted metagenomic analyses have identified genes conferring resistance to β-lactam, tetracycline and glycopeptide antibiotics in a 30000-year-old DNA isolated from the Beringian permafrost [72].

Nowadays, AMR is becoming a concerning global health issue. In 2019, the World Health Organization published a report where it predicted that deaths caused by AMR will reach 10 millions death per year by 2050, surpassing other leading causes of death in humans such as cancer, heart disease and diabetes [73]. The development of AMR is accelerated by the misuse and overuse of antimicrobials and its spread all over the world is facilitated by globalization [74]. In 2008, Yong et al. [75] have identified the mutli-drug resistant gene New Delhi Metallo-β-Lactamase 1 in an Indian patient in Sweden, this illustrates the role of global trade and travel in allowing the resistant organisms to spread further than ever before.

AMR is generating increasing concern for public health worldwide, requiring the implementation of novel strategies for its containment and the mitigation of its dramatic effects.

1.3.2 *In Silico* methods for the characterization of AMR

AMR is serious global healthcare crisis caused by the increasing prevalence of resistant microbes. The identification of successful drug targets allowing the design of efficient therapeutical strategies represent a great challenge for the pharmaceutical industry [76]. In clinical studies, AMR microbes are identified through antimicrobial susceptibility testing by exposing them to different antimicrobials then selecting the microbes which could grow on these lethal conditions. These microbes are isolated and many of their biological characteristics such as genomes and proteomes are made available and stored in database [77]. In the era of Big Data, computational methods can provide effective tools to further analyze these databases in order to complement the conventional experimental protocols of disease characterization and drug design. The use of such methods contribute to lowering the time, the cost and the efforts dedicated to design new therapies [78].

Different *in silico* methods are used to elucidate AMR mechanisms and identify new drug targets. These methods can be classified into three categories basing on

the data that are used: 1) Molecular modeling-based approaches; 2) Machine Learning approaches; and 3) Network-based approaches. Of course, these methods can be combined to complement each other in order to improve the results [79, 80, 81, 82].

- **Molecular modeling-based approaches:** are based on the *in silico* modeling and prediction of drug-target interactions. They require two inputs: 1) the 3D structure of the target of interest (a protein, enzyme...); 2) The 3D molecular structure of the candidate antimicrobial. Proteins' 3D structures can be found in database such as the Protein Data Bank (PDB) [83], if not, homology modeling methods can be used to reconstruct the protein's 3D structure from its amino acid sequence. Antimicrobial's 3D structures, as well as other chemical compounds, can be extracted from chemical libraries like ChEMBL [84], PubChem [85] and DrugBank [86]. One of the methods used for the selection of drug candidates is virtual screening [87]. It relies on an *in silico* binding of the 3D structures of a large chemical library with a protein of interest. The molecule(s) having a high binding affinity are selected as drug candidates for further clinical essays.

- **Machine learning-based approaches:** in the field of drug design, the Quantitative Structure–Activity Relationship approach (QSAR) is widely used [88]. QSAR relies on learning the functional properties of the molecules, such as their antimicrobial and inhibition activities, from their physicochemical features (descriptors) like molecular weight, lipophilicity, number of rings... Then, the QSAR models are used to predict the antimicrobial activity of the new molecules of interest.

  Artificial intelligence (AI) methods have a wider use. In addition to identifying new antibacterials from chemical libraries [89, 90], AI is also used to identify potential targets [91], predict acquired AMR [92] antibiotic susceptibility profiles of microbes [93], recommend antibiotic prescriptions [94] and predict the effect of drugs on biomarkers [95].

- **Network-based approaches:** rely on the analysis of different kinds of networks such as gene networks, protein networks and antibacterial-target networks to detect possible AMR-related genes and AMR-related functional motifs. At least two types of functional motifs [96] can be found in biological networks: 1) Topology-based motifs [97] which are sub-networks, i.e. a set of nodes interconnected together with a certain pattern, e.g. closed triplets; 2) Topology-free motifs [98] are sub-networks of nodes defined not only by the nodes' interconnectedness but also by the nodes' functions. In these kind of motifs, nodes are interconnected and share exactly the same properties (i.e. biological functions), contrary to the topology-based motifs where nodes can have different properties. One topology-free motif can be associated to several functions as biomolecules such as genes and proteins can be pleiotropic[4].

  Carunta et al. [99] could detect functional motifs which can be associated to AMR in *E. coli* from a network of AMR-related genes. By means of centrality measures and community detection approaches, Myryala et al. [100] analyzed the gene network of *P. aeruginosa PA01* strain and could detect hub genes and functional clusters associated to AMR, which could be potential targets to mitigate the multiple drug resistance (MDR) in *P. aeruginosa*. Using a similar approach, Anitha et al. [101] and Miryala et al. [82] have characterized MDR and identified potential drug targets in the pathogens *Acinetobacter baumannii* and *E. coli O157:H7*, respectively.

[4]Pleiotropy: in biology, a gene or protein is said pleotropic if it determines different phenotypic traits.

1.4 Network Science and Constraint-based Modeling

Cells are the basic units of life. Their functions is governed by the interaction of different types of biomolecules such as DNAs, RNAs, proteins and metabolites. These molecules form networks of specific functions: signal transduction networks, gene regulatory networks, protein-protein interaction networks and metabolic networks. These latter play a crucial role in maintaining cell functions as they provide the "fuel" and the "ingredients" indispensable for the functioning of the other biological networks. Therefore, studying metabolic networks can shed light on many unknown biological processes, but before being able to perform such studies, these complex and dynamic networks need to be first modeled, then, analyzed using the appropriate tools. Here Genome Scale Metabolic Modeling (GSMM) and Flux Balance Analysis (FBA) come into play. In the next sections, we introduce GSMM and FBA which we combined with network approaches (Chapter 4) in order to reconstruct context-specific metabolic models of yeast.

1.4.1 Genome Scale Metabolic Modeling

After the complete sequence of the free-living microorganism *Haemophilus influenzae* Rd was made available, the first GSMM was constructed to model the metabolic pathways in *H. influenzae* [102]. GSMMs are built starting from the genome sequence which is annotated with curated biochemical information. The annotated genes define the reactions and their associated metabolites to be included in the model [103]. For a detailed GSMM construction procedures, please check these papers [103, 104, 105]. Here, we describe the four major steps for building a GSMM from a genome :

1. **The construction of the first version of the model from the genomic sequence:** This first version is called a draft GSMM. It contains all the proteins encoded by the genome and their associated metabolites when those proteins are enzymes, and, for the proteins that are transporters, the model accounts also for the metabolites entering or going out of the cell through them. The GSMM is represented as stoichiometric matrix (Table 1.3) where the columns represent the reactions and rows represent the metabolites. The entries can be either positive integers if the metabolite is produced by a reaction(s); negative integers if the metabolite is consumed by a reaction(s); or, a null entry (0) if the metabolite is not associated to the reaction(s) [106]. For instance, reaction R_1 consumes 2 molecules of the metabolite M_1 and 1 molecule of M_2 to produce 1 molecule of of M_3 (Table 1.3).

	R_1	R_2	R_3	R_4
M_1	-2	1	-1	0
M_2	-1	0	1	-1
M_3	1	-1	0	1

Table 1.3: Stoichiometric matrix of a toy GSMM with four reactions and three metabolites.

2. **The refinement of the draft GSMMs by filling the gaps and correcting inconsistencies:** the gene-protein-reaction associations (GPR) are retrieved from databases such as BIGG [107], KEGG [108], BRENDA [109], BioCyc [110] and ModelSEED [111]. These databases may not include all the GPR of the modeled organism, therefore, the draft GSMM should be checked manually to fill its gaps by adding missing reactions. In addition, not all the organisms can be found in these databases. To annotate a new organism's genome, gene sequence similarity is used to define the GPR associations, but sequence similarity doesn't systematically lead to a function transfer [105]. Furthermore, the presence of a gene sequence in the organism's genome doesn't systematically mean that the

gene is expressed and its product is involved in a metabolic pathway. Therefore, in addition to gap filling, draft GSMMs should also be checked for such inconsistencies.

3. **Definition of the biomass equation and the growth medium requirement:** as stated by Santos et al. [112], a GSMM becomes a model only if it includes biomass equation, in other words, GSMMs can't be used for prediction purposes unless they have a mathematical representation. Otherwise, the GSMM is just a reconstruction of the metabolic network.

   The biomass equation (see the example below) determines the chemical composition of the cell. It represents the relative contribution of the cell's organic (proteins, lipids, RNAs, vitamins, cofactors...) and non organic (ions) molecules. The chemical composition can be either determined by experiments, found in the literature, or estimated.

$$\begin{aligned}\text{Biomass} = 0.55 \times \text{Protein} + 0.10 \times \text{RNA} + 0.07 \times \text{DNA} + 0.10 \times \text{Lipids} \\ + 0.09 \times \text{Carbohydrates} + 0.03 \times \text{Cofactors} + 0.01 \times \text{Inorganic ions}\end{aligned}$$

   The illustrative mass equation above determines the relative fractions of the different molecules in the cell.

   Secondly, as energy is required for every biological process, the biomass equation must also account for the energetic costs for maintaining the growth, i.e. how many ATP[5] molecules the cell consumes to ensure its functions such as macromolecular biosynthesis. Like biomass composition, energy requirement can be either estimated or experimentally defined. The following reaction represent the ATP hydrolysis where x is the stoichiometric number of the reaction, i.e. how many reactant/product are consumed/produced:

$$xATP + xH_2O \rightarrow xADP + xP_i + xH^+$$

   Finally, comes the determination of the growth medium requirement, i.e. the essential nutrients that the organism takes up from the medium to ensure its growth, such as vitamins and carbon and nitrogen sources. These data can be retrieved either from the literature or defined through experiments.

4. **Experimental validation of the curated model:** now that the model is available, it should be tested to assess its consistency and predictive performance. To do so, simulation can be done to see wether the cell can grow and reach coherent production of, for instance, CO_2 and ATP. Also, the model can be compared to knock-out cells with a known phenotype. The knocked-out gene can be "switched off" in the model, then a simulation is run. If the model's results are not in accordance with the experimentally observed phenotype, then the model should be checked and curated again to resolve the inconsistency

   Experimental evaluation of the model and manual curation are iterated until obtaining the most consistent model representing as well as possible the metabolic properties of the studied organism.

The reconstruction of good-performance GSMMs is time-demanding, for example, the curation of the model can take months to a year [103]. Therefore, many tools have been developed to expedite the reconstruction of GSMM such as KBase [113], Merlin [114] and ModelSEED [115] (for an extensive review of the available GSMM reconstruction tools, please consult [116, 117]). These open-source tools are available online (e.g. KBase and ModelSEED) or in user-friendly standalone interface (Merlin).

[5]Adenosine triphosphate (ATP): is the molecule used in living beings to generate/store energy.

They allow the user to carry out the different steps required to build GSMMs such as genome annotation, gap filling and validation of the model. Functionalities of these tools are more or less similar, however, they can differ in the way they build a model starting from the genomic sequence: 1) The GSMM can be built from scratch by annotating the different genes and adding the biochemical information extracted from databases (KEGG, BRENDA...); 2) In the case where the new organism to be modeled is phylogenitically close to an organism of which a curated GSMM is already available, one can use a template-based approach, that is, using the information from the existing GSMM to build the new one.

We note that, although GSMMs are comprehensive representation of the cell's metabolic potential, they don't take into account the kinetics of the reactions, i.e. the rate at which each metabolite is consumed or produced. Therefore, GSMMs can't capture dynamic behaviors of the metabolic pathways. Kinetic models, on the other hand, accounts for both the structure and the kinetic details of the reactions. A legitimate question arises: if kinetic models represent metabolism in a more realistic way, why one would use GSMMs? Before answering the question, let's first stress that kinetic models were indeed successfully used in many context to explore and control metabolic pathways, such as in human metabolic disease [118] and metabolic engineering of lactic acid bacteria [119], but such results can be achieved only if a detailed kinetic model is available, i.e, a model that accounts for most (all) the kinetic parameters that govern the biochemical reactions in the system [120]. Such complete models are not often available because kinetic parameters are not available for all the reactions, and, in the rare cases where they can be estimated, the corresponding models are relatively small and isolated from the whole metabolic network of the cell [112]. Given these limitations, GSMMs offer an alternative to cover the metabolic potential of the cell without the need of estimating every kinetic parameter. Of course, discarding these parameters is not totally costless, but, regardless this limitation, GSMMs can be used as an effective tool to explore metabolic pathways and predict their outcomes if combined with adequate methods, such as Flux Balance Analysis (next section).

1.4.2 Flux Balance Analysis

Flux Balance Analysis (FBA) [121] is used to Analyze the metabolites' Fluxes under the assumption that the system is Balanced, in other words, FBA computes the flow of the metabolites in a metabolic network by assuming a steady-state, i.e. a metabolic state where the amounts of produced and consumed metabolites are equal. FBA uses Linear Programming (LP) to maximize or minimize a function of interest, i.e, an objective function. LP finds the optimum of the objective function by computing the parameters that minimize/maximize it. These parameters can be computed from a set of constrained linear equations.

Let's take a simple example to illustrate FBA. Figure 1.3 represent a metabolic network where the substrate S is converted into metabolite M, which is, in turn, converted into two products P_1 and P_2. v_1, v_2 and v_3 are the rates or the fluxes expressed of each of the reactions, i.e. how much of the metabolite is consumed or produced by the reaction per unit time.

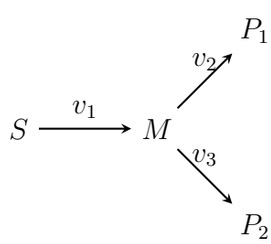


Figure 1.3: Simple reaction pathway involving three reactions, one substrate S, one intermediate metabolite M and two products P_1 and P_2

To run FBA, we have to set the mass balance and the capacity constraints (boundaries) on the fluxes v:

Mass balance:

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

where $\mathbf{S}$ represent the stoichiometry matrix of the metabolic network, and $\mathbf{v}$ is a column vector of the fluxes of each reaction (i.e, v_1, v_2 and v_3). At steady-state, $v_1 - v_2 - v_3 = 0$.

Now, suppose that our objective is to maximize M yield. To do so, we have to maximize the rate of the reaction that is producing M, i.e, v_1. We can formulate the problem as follows:

$$\text{Objective: maximize } v_1$$

$$\text{given:}$$

$$\begin{cases} \text{mass balance constraint:} & v_1 - v_2 - v_3 = 0 \\ \text{capacity constraints:} & -\infty \leq v_1 \leq \infty \\ & 0 \leq v_2 \leq 5 \\ & 2 \leq v_3 \leq 10 \end{cases}$$

Thus

$$v_1 = v_2 + v_3$$

$$max(v_1) = max(v_2) + max(v_3) = 15$$

.

The upper and lower allowable bounds of the different fluxes can be measured experimentally or set according the objective function of the simulation. For example, in our example, if we knew that the cell doesn't produce P_2 in the studied context, we could set $v_3 = 0$.

The cells' metabolic networks are stable and resilient to perturbations due to the redundancy of the enzymes and reaction pathways [122, 123]. Therefore, the optimal state that can be found by FBA is just one state among a set of possible optimal states, as the objective function (e.g. maximization of growth) can be achieved through different reactional routes implying different metabolites fluxes. To capture and measure this metabolic flexibility, one can use a variant of FBA which is Flux Variability Analysis (FVA) [124]. Instead of a unique flux value, FVA computes a range of possible flux values for each reaction.

Furthermore, metabolism is a dynamic and evolving process, thus, the assumption of steady-state is valid only in some particular cases [125]. To cope with this limitation, FBA can be applied sequentially by running it at each time point of the evolution of the system, and this variant of FBA is called Dynamic Flux Balance Analysis (dFBA) [126].

In Chapter 4, we combine FBA/FVA and network-topology-based metrics to create context-specific yeast's GSMM.

Chapter 2

Estimation of metabolite levels in cheese from microbial gene expression

2.0.1 Introduction

Cheese is a dairy product that has been produced since the earliest civilizations some 8000 years ago during the "Agricultural Revolution" [127]. Its production spread throughout Europe and the Middle East and later to North and South America and Oceania. Nowadays, there are at least 1000 cheese varieties produced all over the world [4]. The conversion of fresh milk into cheese involves different microbial species including bacteria, yeasts and moulds which perform the three major pathways constituting the biochemistry of cheese fermentation and ripening: (1) metabolism of residual lactose and of lactate and citrate (primary reactions), (2) lipolysis and fatty acid metabolism (3) proteolysis and amino acid catabolism (secondary reactions). These biochemical processes result in the development of flavor and texture characteristics of the cheese. For instance, Lc. lactis ssp. lactis and Leuconostoc spp metabolize citrate to diacetyl in the presence of a fermentable sugar during manufacture and early ripening. Diacetyl contributes to the flavor of Dutch-type cheeses and possibly Cheddar also. The CO_2 produced is responsible for the small eyes characteristic of Dutch-type cheeses. The metabolism of fatty acids in cheese by Penicillium spp produces n-methyl ketones which dominate the taste and aroma of blue cheese [127].

Cheese flavor and aroma are among the main properties that determine cheese's quality and influence consumers' preferences [12, 13, 14]. For that reason, one of the main issues for the dairy industry is monitoring and characterizing cheese's composition, aroma, flavor and nutritional characteristics during cheese making processes [128]. The most powerful sensory tool in cheese flavor research is descriptive sensory analysis. It requires trained human sensory evaluators to identify and quantify sensory aspects like appearance, aroma, flavor, texture... [129]. A good cheese flavor evaluator requires regular maintenance and 75-100 hours of training, which makes the descriptive analysis of flavor one of the most complex modalities to train [130]. Furthermore, this form of sensory analysis can be challenged with subjectivity on the side of less trained or unprofessional panelists, the sensitivities of smell receptors [131] and taste buds [132], since the sense of smell and taste varies with age, and in some cases, sex [133, 131] and lifestyle activities such as smoking [134]. Thus, human evaluation and inspection during food quality control may lead to inconsistent decisions. To cope with these challenges, metabolomics technology offers different tools that can identify and quantify the cheese's flavors and aroma such as gas chromatography, mass spectrometry, aroma extract dilution analysis (AEDA) and odor activity value (OAV) [135]. Gas chromatographic methods are widely applied in food science and technology due to an efficient compound separation and versatility. However, these metabolomics analyses demand expensive instrumentation and are time consuming as the optimum recovery of flavor compounds usually require more than one procedure to avoid degradation and formation of artifacts and reach a detectable concentration of the components [136, 130, 137].

The final characteristics of a cheese, mainly flavors and aroma, are due to the complex dynamics and biochemical reactions involving the enzymes produced by cheese microorganisms and the milk components (lactose, fats, proteins). Therefore, analyzing these enzymatic profiles through sequencing methods such as metagenomics, proteomics and transcriptomics can serve to predict and characterize the metabolic profile of cheese. These sequencing methods are well developed and relatively cost-effective compared to the metabolomics experiments that directly measure the metabolic profiles of dairy products [138]. Using an integrative approach to analyze 16S rRNA sequencing and metabolomics data collected from industrial and artisanal cheddar cheeses, Ashfari et al [15] could detect strong relationships between the cheese microbiota and metabolome and uncovered specific taxa and metabolites that contributed to these relationships. Bertuzzi et al [17] have investigated the metabolic potential of the resident microorganisms of a surface-ripened cheese through whole-metagenome shotgun sequencing. They showed how variations in the microbial populations influence important aspects of cheese ripening, especially flavor development. In human-based studies, Mallick et al [139] have developed MelonnPan, an elastic net regularization method that can predict gut metabolites from the metagenomic data of gut microbial communities. This approach displayed promising performance and can be used for the prediction of metabolomes in similar studies where only microbiome is available. Similarly, neural networks have been developed to infer metabolic profiles from metagenomic and uncover microbe–metabolite relationships in environmental and clinical settings, such as soil biocrust wetting, cystic fibrosis and inflammatory bowel disease [140, 141, 142].

Objective of the work

To our knowledge, the possibility of inferring cheese metabolic profile from microbial metatranscriptomics has not been sufficiently explored yet. In this work, we investigated the metatranscriptome-metabolome relationship in an experimental surface-ripened cheese by means of predictive models and correlation analyses. We trained Elastic Net and Random Forest (RF) models to infer the metabolic outcome of the cheese from its microbial gene expression profile. Ultimately, cheese quality control (monitoring metabolic/flavor profile) could be complemented by such straightforward *in silico* predictions, which are more cost-effective and less time consuming than the traditional sensory analyses techniques. This can reduce cheese making costs, especially for large-scale cheese/food manufacturers that are very aware of customer preferences in terms of quality and affordability (cost).

Summary

Despite the spareness of the data, the accuracy of the models could reach 50 to 83 %. Moreover, the analysis of the genes selected by the modeling procedure showed their consistency with biological pathways, mainly metabolic ones. Our results demonstrate that metatranscriptomics data can be used as an informative proxy to estimate flavor profiles in cheese.

2.0.2 Materials and Methods

In this section, we describe the different steps we carried out to build the predictive models and investigate their biological relevance. In a nutshell, these steps included: 1) Pre-processing and transformation of the raw data; 2) Feature selection, construction and assessment of the predictive models; 3) Investigation of the models signatures's correlation with the microbes and biological pathways.

Note: metatranscriptomics and metametabolomics stand for the transcriptomics (gene expression) and metabolomics data, respectively, associated to microbial species in a given environment. For the sake of clarity, we use the terms transcriptomics and metabolomics to refer to these data from now on. In addition, we write the names of the predictive models with a capital letter, e.g, Alcohols is the model that predicts alcohols amounts.

transcriptomics and metabolomics data used for training and testing the predictive models

Our data consist of longitudinal transcriptomics and metabolomics data collected from two independent experiments, that were used separately for training and testing our classification. Each of the training and test datasets contain one transcriptomics (the predictors) set and one metabolomics set (the outcomes). The rows in the trancriptomics datasets contains the time points and the columns contain the sequenced genes. In the metabolomics data, rows represent the same time points and columns contain 6 measured metabolite classes: Alcohols, Aldehydes, Alkanes, Ketones, Esters and Sulphur compounds.

The training data were collected from an experiment with a surface-ripened cheese, fermented and ripened using a reduced microbial community composed of six bacteria and three yeasts: *Glutamicibacter arilaitensis* (Ga), *Brevibacterium aurantiacum* (Ba), *Corynebacterium casei* (Cc), *Hafnia alvei* (Ha), *Lactococcus lactis* (Ll), *Staphylococcus equorum* (Se), *Debaryomyces hansenii* (Dh), *Geotrichum candidum* (Gc) and *Kluyveromyces lactis* (Kl). There were three different experimental conditions: Control, and two other conditions where the Dh or the Gc yeast was omitted from the medium. There were three replicates at each of the time points Day 7, Day 14, Day 24 and Day 31.

The test data set, as provided by Dugat-Bony et al. 2015 [143], is identical to the training one in terms of cheese type and microbial community, except for the experimental conditions (perturbation) which were set according to a variation in NaCl concentration. There were three replicates at each of the time points Day 1, Day 7, Day 14, Day 24 and Day 31.

Normalization of the transcriptomics data

To be able to use transcriptomics data for any analysis, one should consider eliminating, or, at least, reducing the biases induced by biological and technical artifacts as they can create a spurious variability between the studied samples. Here, we introduce two biological and one technical artifacts we should deal with our data:

- **The sequencing depth:** is a common technical artifact in sequencing. It represents the resolution of the sequencing, i.e, how many times the gene sequence is read during the sequencing. Sequencing depth varies between samples and this results in a variation of the total read counts per sample [144]. For instance, when the sequencing depth is two times higher in sample 2 than sample 1, the genes read counts will be duplicated and the gene expression will seem increased while it's actually not.

- **Microbial abundance:** is a biological artifact encountered with microbial data due their evolving behavior (growth or decay over time). This makes the genes expression seem to be varying (increasing or decreasing) across samples.

  Suppose we measured the expression of one gene in one microbial community in two conditions, control and treatment, to see whether the treatment has an effect on the expression of the studied gene. Suppose we found that the gene was ten times more expressed in the treatment condition, and at the same time, growth data showed us that the microbial community has also grown by ten times. In this case, the variation in gene expression doesn't reflect a real shift in the biological behavior of the microbial community under the treatment, as the number of expressed genes increased just because there were more expressing microbes.

- **RNA composition:** is the second biological artifact we should reduce. To illustrate the concept, suppose this time we are interested in the expression of gene A in one microbe, in control and treatment conditions. Suppose we measured 10 reads of gene A in the control and only 2 reads in treatment condition. Could we say that the treatment decreased the expression of gene A? Yes, it could

be. But, this decrease could be also due to a high expression of another gene B which competed for the available sequencing machinery (polymerases, primers, nucleotides...) available in the sequencing pool. As a results, gene A will be less accessible for the sequencing machinery mostly saturated by by gene B, and thus, gene A will seem under-expressed by the microbe while it was just not enough sequenced as in the control condition.

To deal with these biases, we adopted a method proposed by Klingenberg et al. [145]. The procedure relies on splitting the transcriptomics data by microbial species then performing a normalization method that accounts for the sequencing depth and the RNA composition, such as the Trimmed Mean of the M-values normalization (TMM) [146], on each subset.

Both training and test transcriptomics data were split into nine species-specific datasets, each one representing a given microbial species. In each species-specific dataset the lowly expressed genes, i.e, genes with less than 10 reads in three randomly selected samples, were filtered out. After this step, only two genes remained in *Staphylococcus equorum* data in the test set, thus, this species was discarded. As a result, we got nine and eight filtered species-specific datasets in the training and test set, respectively. Then, these datasets were normalized using TMM normalization, implemented in *edgeR* R package [147], to reduces the variations between samples due to the sequencing depth and the RNA composition.

This procedure allows to reduce the variation in gene expression profiles across samples which is due to a variation in microbial abundances and sequencing artifacts, and not to an actual change in functional and biological profiles. After this processing, the remaining gene expression differences reflect better the different behavior of organisms under the changing conditions. Finally, the filtered and normalized species-specific datasets were merged to re-obtain the entire training and test datasets used for the modeling procedure.

Orthology-based shrinkage of transcriptomics data

The metabolic profile in cheese is the result of the contribution of different enzymes expressed by the different microbes. Despite their differences, bacteria and yeasts can express proteins and enzymes having similar functions. These proteins (genes) are called orthologous proteins/genes [148]. To reduce the data's dimensionality, we chose to sum together the expressions of the orthologous genes (same KEGG Orthology identifier [149]). This allowed the shrinkage of the sample sizes in both training and test sets to almost the half (Table 2.1). This is a first dimensionality reduction step which is required before training predictive models. The second step is features selection which is described farther in this chapter.

	Raw	Lowly expressed genes filtered	Orthologous genes merged
Training set	9868	8210	4245
Test set	3899	1769	1279

Table 2.1: Sample size of the training and test datasets before and after filtering out the lowly expressed genes and merging of the orthologous genes. The number of observations is 34 and 24 in the training and test set, respectively.

Metabolomics data processing

Alongside gene expression, the amounts of six metabolite classes were measured in the two experiments: Alcohols, Aldehydes, Alkanes, Ketones, Esters, and Sulphur compounds. So, for each of the training and test sets, we had the corresponding metabolites amounts (observations) per each time point. The original metabolomics data are numerical, so we first trained regression models to estimate metabolites numerical amounts from transcriptomics, while bearing in mind that regression is not the best choice as the observations are very sparse (34 observations). We trained Elastic Net (EN) [150] and generalized additive models (GAMs) [151] after log and

Z-transformation of the data. Indeed, the predictive performances were very poor (low R^2[1], Table 2.2).

Metabolite	Elastic Net	GAM (log transformed)	GAM (z-transformed)
Alcohol	0.0034	0.075	0.114
Aldehyde	0.0253	0.038	0.149
Alkane	0.0415	0.0033	0.062
Ketone	0.0172	0.041	0.009
Ester	0.0330	0.018	0.004
Sulphur compounds	0.0008	0.072	0.006

Table 2.2: R-Squared (R^2) values of the trained EN and GAM regression models.

To cope with this problem, we decided to train classification instead of regression models as the prediction of a class is "easier" than predicting an accurate continuous value. As expected, the models' performances improved (see Results section).

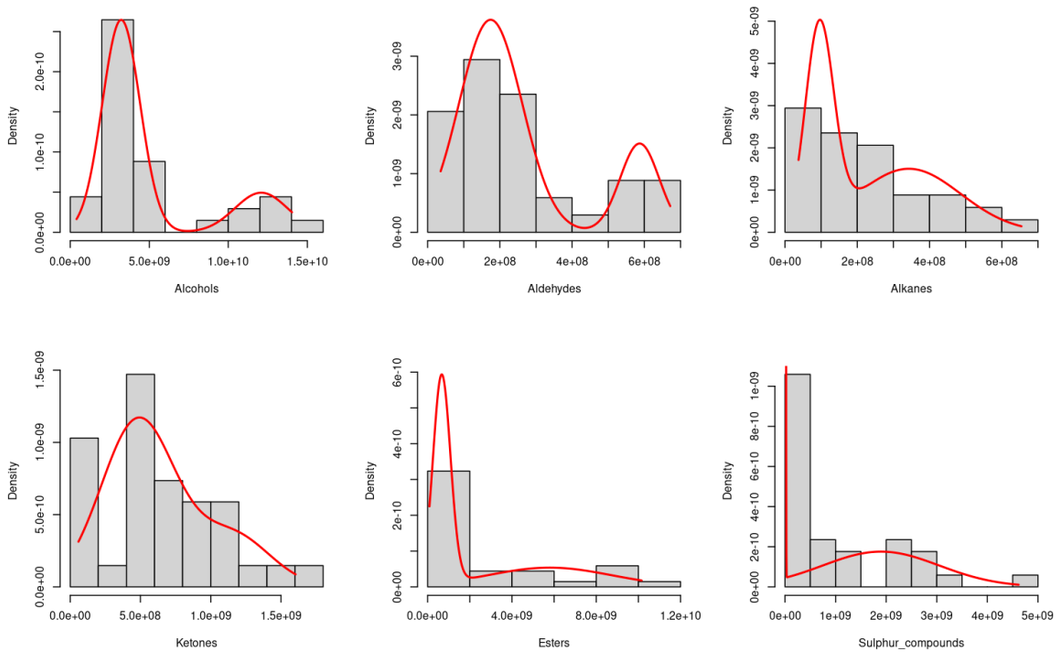
To do so, we converted the numerical values into levels (categories). Given that there were no available human labeling of each observed amount, we investigated the metabolites distribution in order to find patterns allowing classifying them into categories.

The metabolite distributions approximately follow a bimodal distribution (Figure 2.1). We run the Expectation-Maximization (EM) algorithm[2] [153] to try to fit two gaussians to these data. To set the priors of the EM algorithm, we run k-means clustering[3] [155] with two centers ($k = 2$). Then, for each cluster determined by k-means, we computed its mean, standard deviation and proportion. These values were then updated by the EM algorithm, and after convergence, we obtained the estimated means, standard deviations and proportions. These values represent the posteriors we used to model the density of the gaussians (red line in figure 2.1).

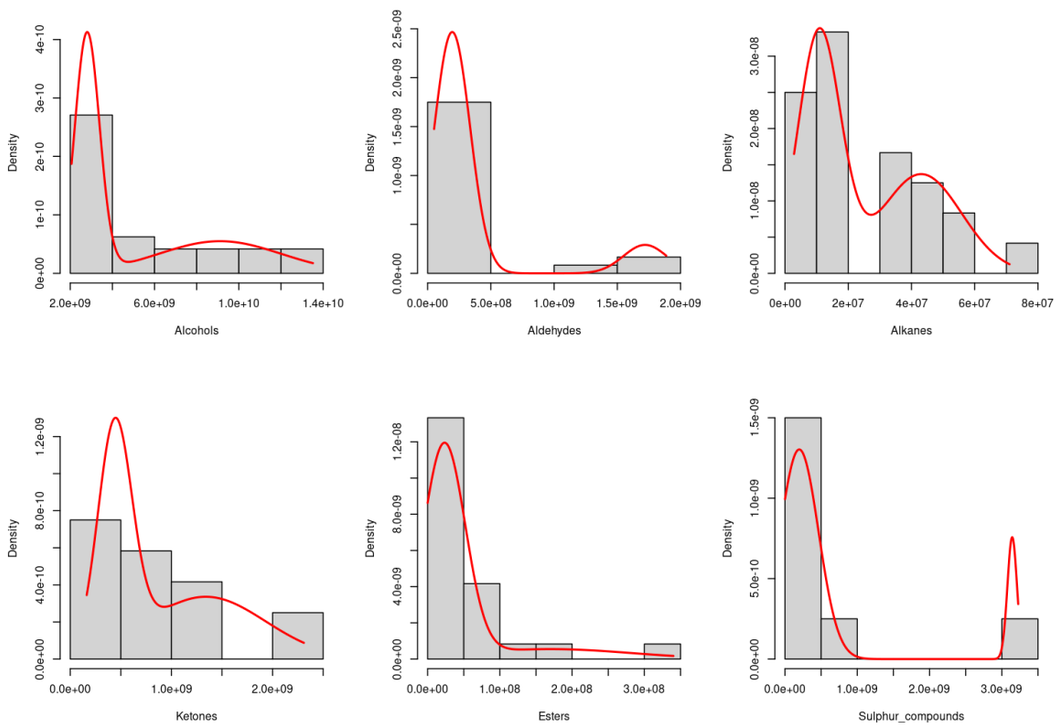
[1]R-Squared (R^2): also called coefficient of determination. It can take values between 0 and 1. It represents how well the genes (predictors) explains the variation of the actual metabolites amounts.

[2]normalmixEM() function from the *mixtools* R package [152]

[3]kmeans() function R version 4.2.2 [154]



(a)



(b)

Figure 2.1: Distribution of the amount of each metabolite in the training (a) and test data (b). Red line: fitted gaussian using Expectation-Maximization algorithm

We labeled the continuous values in each cluster by “High” or “Low”. Note that this labeling is relative and arbitrary. It’s used only to distinguish the two clusters, it could be ”Level 1” and ”Level 2”, ”Class1” and ”Class2”... This labeling is purely statistical, in other words, it doesn’t necessarily correspond to what a cheese expert would qualify as ”High” or ”Low”.

Each of the two data sets as well as the metabolite class were clustered and labeled separately. The original values of the metabolites amounts and the created labels are provided in the supplementary table 2.14 (Only for the test datasets. The training datasets are not published yet).

Training data and test data: further processing

In this section, we give more details about the processed training and test data sets. We describe the issues we had to deal with before being able to use these data for training and validation the predictive models.

1) Same metabolites but different genes

The two transcriptomics datasets were collected from two similar but not totally identical experiments. The expression profiles of microbes depend on the conditions in which they grow, therefore, the number as well as the type of the expressed genes vary from a time point to another, and, from a condition to another.

Given that training and test transcriptomics data were collected using two different experimental designs, the genes found in these datasets don't overlap totally. In the training set 4245 expressed genes were detected while in the test set there were 1279. The number of common genes between those dataset is 1062 genes (based on KEGG [149] Orthology Numbers). The problem that arises here is that we can't train a model on a given set of features then validate it in a different set, even if the predicted variable (metabolite class) is exactly the same. So we had two options: 1) training the model using the full training set (FTS, 4245 genes and 34 observations) and cross-validate them using the same dataset, i.e. FTS; 2) Training the models on a reduced training set (RTS) which is a subset of FTS containing only the 1062 common genes with the test set, then, validate the models on the test set (Table 2.3). Option 1 is good in terms of completeness, i.e, all the microbial profile is explored by the modeling procedure to find the best fitting representing the transcriptome-metabolome relationship, but, on the other hand, the validation is less robust as the models are cross-validated. Option 2 offers a robust validation, as the test set is independent from the training one, but, almost half of the predictive features (genes) are not explored, and this may result in a less accurate estimation of the transcriptome-metabolome correlation in cheese.

To guarantee both completeness and robustness, we performed two modeling procedures in parallel: 1) We trained and cross-validated the classification models using FTS consisting of 4245 genes/34 samples and, 2) We trained the models using RTS (1062 common genes/34 samples) then validated them on the independent test set.

	Training Set (FTS)	Test Set	Training Set ∩ Test Set	Reduced Training Set
Genes	4245	1279	1062	1062
Samples	34	24	-	34

Table 2.3: The size of the original training (FTS) and test sets, in addition to the created Reduced Training Set.

2) The data are cursed and imbalanced!

Let's remind that both FTS and RTS are dimensionally high, i.e, there are more predictors (genes) than observations (levels of metabolites). There are 34 data points (metabolite levels) in FTS and RTS, and 4145 and 1062 genes, respectively. This induces what is called "The Curse of Dimensionality" [156] which is problematic because:

- At a fixed number of data point, the performance of a predictive model increases as the number of predictors increase until it reaches an optimum corresponding to the optimal number of predictors. Exceeding this optimum results in a decrease of the model's performance. This is known as the Hughes Phenomenon [157].
- Few data point with a high number of predictors cause the model to overfit, i.e, it can perfectly fit the patterns in the training data but fails to generalize, i.e, perform good predictions from new data.
- Too many features increases the number of redundant and correlated predictors. A model with many correlated predictors is less informative and generalizes poorly.

Secondly, we investigated the levels of the six metabolite classes and we found a class imbalance, i.e, the "High" label is much more lower than the "Low" one Table 2.4. When the response variable (classes) is skewed, it can break down relatively robust procedures used for unskewed data [158].

To handle the high dimensionality and class imbalance, we performed and compared three classification approaches (described below): 1) EN models to handle multicollinearity and shrink the data; 2) Random Forest classifiers to handle class imbalance; 3) EN-based feature selection followed by RF classifiers to cope with both problems.

Metabolite	Training set	Test set
Alcohols	7/27	6/18
Aldehydes	7/27	3/21
Alkanes	12/22	10/14
Ketones	12/22	8/16
Esters	9/25	2/22
Sulphur compounds	9/25	3/21

Table 2.4: The number of High/Low labels associated to each numerical value of the metabolite amounts.

Elastic Net classifiers to handle multicollinearity

As stated earlier, we wanted to train classification models that predict the Low/High levels of a given metabolite class, based on microbial gene expression values. More precisely, we trained six different models, each one predicting one metabolite class. In a similar work, Mallick et al. [139] have trained elastic net regularization model to predict gut metabolites from the metagenomics data of human gut microbial communities. In our work, we adopted a similar methodology.

To handle high dimensionality, there exist shrinkage methods such as Ridge [159] and LASSO [160]. These two methods are designed to shrink the solution space by constraining the model's coefficients. The regularization of the model is needed in the case of high dimensional data to get more stable models. Ridge was initially designed to handle correlated predictors. It can assigns very small coefficients to non-relevant predictors, but never assigns a null (0) coefficient. LASSO, on the other hand, can shrink the number of the predictors by assigning a null coefficient. EN combines Ridge and LASSO features, i.e, it handles multicollinearity (predictors correlation) and assigns null coefficient to non informative predictors [161, 162], in addition, EN can be used for both regression and classification.

In regression models, EN cost function combines Ridge and LASSO as follows:

$$J(\theta)_{\text{ElasticNetRegression}} = \underbrace{\frac{1}{2m}\sum_{i=1}^{m}(h(x_i)-y_i)^2}_{\text{Mean Squared Error (Data fitting term)}} + \underbrace{\lambda\sum_{j=1}^{n}\left[\frac{1}{2}(1-\alpha)\theta_j^2+\alpha|\theta_j|\right]}_{\text{Regularization term (L1 + L2)}}$$

with θ are the model's coefficients, m is the number of the observations (34 in our case), $h(x_i)$ and y_i are the predicted and true values of x_i, respectively. λ is a tuning parameter that controls the overall strength of the penalty imposed by the Regularization term [150, 163]. We can see how the Regularization term in EN combines Ridge (L2 penalty: θ_j^2) and LASSO (L1 penalty: $|\theta_j|$) by means of the parameter α which is used to set the compromise between Ridge and LASSO:

$$0 < \alpha < 1 : \text{compromise between Ridge and LASSO}$$

$$\alpha = 0 : \text{Ridge regularization}$$

$$\alpha = 1 : \text{LASSO regularization}$$

As stated earlier, we chose to build classification models in our study to predict the level of the metabolites ("High/Low") instead of continuous numerical values. To use EN for classification, the Mean Squared Error term is replaced by the Logistic Loss term as follows:

$$J(\theta)_{\text{ElasticNetClassification}} = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left[-y_i \log(h(x_i)) - (1-y_i)\log(1-h(x_i))\right]}_{\text{Logistic Loss (Cross-Entropy)}} +$$

$$\lambda \underbrace{\sum_{j=1}^{n}\left[\frac{1}{2}(1-\alpha)\theta_j^2 + \alpha|\theta_j|\right]}_{\text{Elastic Net Regularization (L1 + L2)}}$$

with:

$$h(x_i) = \frac{1}{1+e^{-\theta^T x_i}}$$

To train the EN classifiers, we used the *caret* [164] and *glmnet* [163] R packages.

Random Forest to handle class imbalance

The examination of the response variable (Low/High classes) related to the training and test data sets revealed a class imbalance for most of the metabolite classes, where the "Low" class is bigger than the "High" one (Table 2.4). In a similar context, Liu et al. [165] compared Random Forest, Support Vector Machine and back propagation neural network in classifying imbalanced metabolomics data and they showed that Random Forest was the best to cope with imbalanced learning.

The idea behind Random Forest (RF) [166] is that, instead of trying to build one powerful model, one can build many independent models (the trees) of which the votes are combined (the forest) to make the final decision. These kind of machine learning methods that aggregate many predictive models are called ensemble methods.

To give a real-life example that illustrates the idea behind RF, suppose you have some symptoms (predictors) and you want to know whether you have a serious or benign health problem (i.e, the class to predict: serious/benign). Instead of seeing one physician (model), whose diagnostic can be either judicious or wrong, it's better to see many physicians (independent models). The final diagnostic about your health state will be the result of aggregating the votes of the different physicians.

Now that we know where the the word "Forest" comes from, the adjective "Random" describes how the decision trees are built withing the forest. To make sure that every tree is different than the others, and thus, yields a quite different information, RF uses two random sampling methods: bootstrap and random feature sampling.

To illustrate that, suppose the rows in our training data represent the observations (High/Low metabolite level) and columns represent the predictors (the genes): 1) Bootstrap randomly samples with replacement rows (the observation) from the training data; 2) Random feature sampling randomly select a set of predictors (columns); 3) The randomly selected observations and predictors now constitute a sub-training set on which a decision tree is built. Finally, the obtained trees are aggregated to build the Forest. This step is called Bootstrap Aggregation or Bagging. The Forest is then used to predict the outcomes from new data. For the details and the steps of decision tree construction, please refer to and for intuitive tutorials and illustrations [167, 168, 169].

Bootstrap helps to reduce the data imbalance as the random selection of observations can result in sub-training data sets where the classes "High" and "Low" are more or less equally represented. On the other hand, random feature sampling contributes to the improvement of RF models as they reduce the total variance of the created Random Forest (V_{forest}). The fact of building trees from different randomly selected predictors results in a less correlation between these trees:

$$V_{\text{forest}} = \rho\sigma^2 + \frac{1-\rho}{T}\sigma^2$$

with T is the number of the trees in the Random Forest, ρ is the average correlation between them and σ^2 is their variance. We can see how a reduced ρ and a bigger T can decrease the total variance of the Random Forest.

In our second modeling approach, we trained RF classifiers using *caret*v 6.0.94 [164] on the original FTS and RTS, without prior selection of predictors. We have seen that the performances of two different RF models can be significantly different. Therefore, to effectively assess our modeling procedure, we run 1000 replications with 1000 different random seeds. As a result, we got 1000 similar but not identical RF models then we averaged their performances to get an overall estimation (Table 2.8).

EN feature selection and RF

In our third modeling approach, we wanted to deal with both high dimensionality and class imbalance by: 1) Performing a feature selection, i.e, a shrinkage of the data using EN; 2) Training of RF classifiers using the selected predictors (see Results).

Biological validation of the selected features

To assess the biological relevance of the selected genes (predictors), we performed a Spearman correlation test [170] between their eigengenes and the amounts of each metabolite class. The eigengene E is the first principal component summarizing the expression profiles in each selected gene set [171]. Then, we performed a permutation test to see how significant is the EN-feature selection as compared to a random gene selection.

Secondly, these genes have been analyzed to assess their relevance and association to biological pathways through over-representation analysis (ORA), using the *enrichKEGG()* function from *clusterProfiler* R package [172]. The expressed genes have been chosen as background gene list (gene universe). As metabolic pathways are interconnected and often overlap [173], i.e, enzymes and metabolites of a specific reaction chain can be also involved in another reaction chain, we merged selected predictors of the six models to construct the gene set used to run the ORA.

Assessment of the classification models

For a more accurate estimation of the RF models' performances, we used metrics suitable for imbalanced learning: Balanced Accuracy, Area Under ROC curve (AUC), F1 score and Specificity.

First of all, we define the "Low" class as the Positive (P) class/event, and the "High" as the Negative one (N). True Positives (TP) and True Negatives (TN) are the correctly predicted "Low" and "High" classes, respectively. False Positives (FP) and False Negatives (FN) are the incorrectly predicted "Low" and "High" classes, respectively.

Accuracy is the simplest metric used for the evaluation of predictive models. It is the proportion of good predictions,, i.e, TP and TN out of all the predictions (TP+FP+TN+FN). Accuracy is widely used, however it can be misleading in the case of imbalanced classes.

Therefore, we used *Balanced Accuracy* as an alternative as it takes into account both the proportions of correctly predicted positive (Sensitivity or recall) and negative classes, then averages them:

$$\text{Balanced Accuracy} = \frac{\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}}{2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Area Under the Receiver Operating Characteristic Curve (AUC) measures how well the model discriminate between "Low" and "High" classes. $0.5 \leq AUC \leq 0.6$ indicates that the models prediction are random. $AUC > 0.7$ indicates that the model has fair to excellent (AUC = 1) discrimination ability [174]. The ROC curve plots the variation of the sensitivity of the model (also called *Recall* or *True Positive Rate (TPR)*) as a function of the variation of the False Positive Rate (FPR, i.e, 1-Specificity). The AUC is computed as follows:

$$\text{AUC} = \int_0^1 \text{TPR}(FPR)\, d(\text{FPR})$$

F1-measure represent the harmonic mean of *Precision* and *Recall*:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

with

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The models trained on the FTS were assessed using a Leave-One-Out cross-validation (LOO-CV), whereas those trained on the RTS were estimated using the independent test set. In the case of RF models, the performances of the 1000 replicates were averaged to get an overall estimation. Feature selection and model fitting were performed using the R package *caret*v 6.0.94 [164].

Analysis of microbes' contribution to the selected features

The selected gene sets from both FTS and RTS were analyzed to define which gene(s) is expressed by which microbe(s). Each organism's contribution is represented by a percentage:

$$\text{Contribution percentage} = \frac{\text{the number of the microbe's genes in the selected set}}{\text{total number of the microbe's expressed genes}}$$

Microbes highly involved in metabolic process leading to flavor formation should have high contribution percentages, i.e, the feature selection step should capture more genes expressed by those microbes. To check this assumption, we measured the relationship between the different microbial species and the metabolites classes. We computed a Spearman correlation [170] between the growth rate of these microbes (CFUs[4]) and the measured amounts of metabolites across the samples.

2.0.3 Results

EN performances

In the Methods section, we showed that EN regression models performed poorly in predicting the metabolite amounts (numerical values) from the transcriptomics data. Here, we present the performance of the EN classifiers trained on both FTS and RTS, then cross-validated using LOO and validated using the independent test set (Tables 2.5 and 2.6, respectively).

The cross-validation results are overall very good, except the specificity values. We can see that Alcohols and Aldehydes models could correctly predict only around 50 % of the minority class "High".

[4]Colony Forming Units: is a unit used to measure microbial growth.

Model	Balanced Accuracy	AUC	F1	Specificity
Alcohols	0.71	0.88	0.93	0.43
Aldehydes	0.71	0.88	0.93	0.43
Alkanes	0.79	0.80	0.90	0.58
Ketones	0.83	0.83	0.92	0.67
Esters	0.94	0.98	0.98	0.89
Sulphur compounds	0.94	0.97	0.98	0.89

Table 2.5: Performances of the EN classifiers trained on FTS and cross-validated using LOO.

Model	Balanced Accuracy	AUC	F1	Specificity
Alcohols	0.47	0.79	0.83	0
Aldehydes	0.48	0.68	0.91	0
Alkanes	0.48	0.51	0.69	0.1
Ketones	0.47	0.38	0.52	0.50
Esters	0.70	0.93	0.93	0.50
Sulphur compounds	0.38	0.30	0.16	0.67

Table 2.6: Performances of the EN classifiers trained on RTS and validated using the independent test set.

The EN models trained on the RTS and validated on the independent test set yielded poor to moderate performances (BA < 0.48 and Specificity $\in [0, 0.67]$). Esters is the only model with satisfying performances (BA $= 0.70$ and Specificity $= 0.50$).

The class imbalance can be the cause of the poor performance of these classifiers, especially, in predicting the less represented class "High", which is reflected by their low specificity scores. To try to improve the specificity of the models, we used Random Forest classifiers (next section).

RF performances

This second classification approach aimed at improving the classification performances by using RF classifiers which can cope with imbalanced data.

RF models trained on FTS have more or less similar performances as the EN counterpart (Table 2.7)

Model	Balanced Accuracy	AUC	F1	Specificity
Alcohols	0.68	0.90	0.90	0.43
Aldehydes	0.68	0.90	0.90	0.43
Alkanes	0.80	0.76	0.89	0.63
Ketones	0.76	0.75	0.85	0.61
Esters	0.94	0.99	0.98	0.89
Sulphur compounds	0.84	0.95	0.92	0.75

Table 2.7: Performances of the RF classifiers trained on FTS and cross-validated using LOO.

As compared to EN classifiers trained on RTS (previous section), RF models overall resulted in a net improvement of the prediction performances (Table 2.8), except for the Sulphur compounds model which totally failed to predict the minority class "High" (Specificity $= 0$).

Model	Balanced Accuracy	AUC	F1	Specificity
Alcohols	0.77	0.89	0.87	0.69
Aldehydes	0.83	0.96	0.87	0.87
Alkanes	0.56	0.63	0.50	0.73
Ketones	0.44	0.51	0.60	0.31
Esters	0.50	0.90	0.09	1
Sulphur compounds	0.50	0.60	0.93	0

Table 2.8: Performances of the RF classifiers trained on RTS and validated using the independent test set.

EN+RF performances

In the third modeling procedure, we combined EN and RF to: 1) shrink the data set and select the most relevant predictors; 2) train RF classifiers. We were aiming at handling both predictor multicollinearity and class imbalance. Here we show the RF performances obtained by both cross-validation (LOO) and validation using the independent test set.

All the models performed very well in terms of accuracy (BA $\in [0.82, 0.94]$) as well as in prediction of the less abundant class "High" (Specificity $\in [0.67, 0.89]$) (Table 2.9). These are promising results however, the cross-validation is not so robust as the models were trained on FTS and validate on it using LOO.

Model	Balanced Accuracy	AUC	F1	Specificity
Alcohols	0.83	0.96	0.94	0.72
Aldehydes	0.83	0.96	0.94	0.72
Alkanes	0.82	0.87	0.90	0.67
Ketones	0.82	0.92	0.90	0.67
Esters	0.94	0.99	0.98	0.89
Sulphur compounds	0.91	0.98	0.96	0.86

Table 2.9: Performances of the RF classifiers trained after EN-feature selection, on FTS and cross-validated using LOO.

For a more robust estimation of the models' performances, the RF were trained using the predictors selected from the RTS then validated on the independent test set. In terms of accuracy, all models were rather good (BA $\in [0.49, 0.83]$). Alcohols, Aldehydes and Esters models had good specificity scores, 0.76, 0.77 and 1, respectively, while the specificity score for the remaining models ranged from 0 to 0.40. The Sulphur compounds model still failed to predict any "High" class (Table 2.10).

Model	Balanced Accuracy	AUC	F1	Specificity
Alcohols	0.83	0.92	0.91	0.76
Aldehydes	0.79	0.84	0.87	0.77
Alkanes	0.49	0.54	0.58	0.40
Ketones	0.49	0.52	0.79	0.007
Esters	0.58	0.92	0.27	1
Sulphur compounds	0.50	0.44	0.93	0

Table 2.10: Performances of the RF classifiers trained after EN-feature selection, on RTS and validated using the independent test set.

Correlation between the selected predictors and the biological traits

The EN features selection has retained about 0.004-0.03 % and 0.007-0.08 % of the total number of genes in FTS and RTS, respectively. To assess the correlation between the selected predictors and the metabolites classes, each predictors set has been represented by its eigengene E (the first component vector), then Spearman correlation has been computed between E and the corresponding metabolite class (Table 2.11). In both

FTS and RTS, E correlates well with the corresponding metabolite class where the highest correlation was observed for Sulphur compounds and the lowest for Alcohols.

Despite the strong correlations between aldehydes and esters and the corresponding selected predictors, they are not significantly higher than correlations obtained by a random predictor selection. In the case of Aldehydes, 20 % and 15 % of the random correlations were equal or higher than 0.66 and 0.68, respectively. For Esters, 23 % and 42 % of the permuted correlations were equal or higher than 0.85 and 0.77, respectively. This could be due to the high dimensionality of the data which causes the model to be unstable, i.e, it can find several patterns that explain well the metabolite-genes relationship.

Model	# Selected genes from FTS	Correlation with the metabolite	# Selected genes from RTS	Correlation with the metabolite
Alcohols	45	0.57	23	0.51
Aldehydes	45	0.66*	23	0.68*
Alkanes	17	0.79	7	0.76
Ketones	48	0.79	28	0.74
Esters	128	0.85*	82	0.77*
Sulphur compounds	42	0.92	25	0.92

Table 2.11: Number of features selected in each modeling procedure and the correlation of their eigengene with the metabolites classes. *: the correlation is not better than random (pvalue > 0.05).

Secondly, to assess the biological relevance of the selected predictors from both FTS and RTS, we run an over-representation analysis to see which biological pathways were significantly represented. Seven metabolic pathways were associated to the predictors selected by the EN algorithm (Table 2.12).

Microbial metabolism in diverse environments, Phenylalanine metabolism, Biosynthesis of secondary metabolites, Carbon metabolism, Citrate cycle (TCA cycle) and Glyoxylate and dicarboxylate metabolism. Biosynthesis of secondary metabolites was exclusively enriched in FTS whereas Pyruvate metabolism was exclusively enriched in RTS (Table 2.12). Cheese ripening, mainly, the development of aromatic compounds (alcohols, aldehydes, ketones...) involve many metabolic pathways such as carbon metabolism and citrate cycle [10, 175].

Over-represented Pathways	**FTS**	**RTS**
Microbial metabolism in diverse environments	✓	✓
Phenylalanine metabolism	✓	✓
Biosynthesis of secondary metabolites	✓	
Carbon metabolism	✓	✓
Citrate cycle (TCA cycle)	✓	✓
Glyoxylate and dicarboxylate metabolism	✓	
Pyruvate metabolism		✓

Table 2.12: KEGG pathways over-represented using the genes selected from FTS and RTS.

The selected genes will analyzed individually by the biologists with whom we collaborate. In Table 2.13 we report some of the genes encoding enzymes involved in metabolic pathways. The entire selected gene set will be analyzed to depict the interconnections between the metabolic pathways in which these genes are involved.

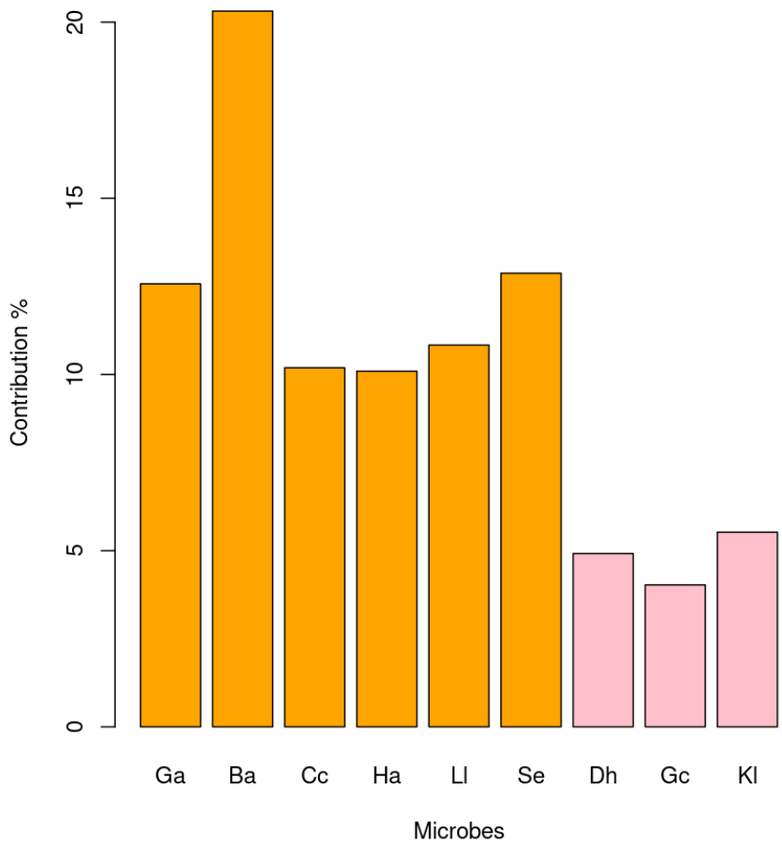
KEGG ID	**Name**	**Function**
K04021	aldehyde dehydrogenase	conversion of aldehydes into acetates
K00121	alcohol dehydrogenase	conversion of primary and secondary alcohols to the corresponding aldehyde or ketone
K01579	aspartate 1-decarboxylase	synthesis of vitamin B_5 required to synthesize coenzyme A which is essential for cellular energy production for the synthesis and degradation of proteins, carbohydrates, and fats.
K00232	acyl-CoA oxidase	enhances ketone formation
K00683	glutaminyl-peptide cyclotransferase	degradation of peptides into free amino acids
K01069	hydroxyacylglutathione hydrolase	D-lactate biosynthesis from methylglyoxal

Table 2.13: Non-exhaustive list of the genes selected by the EN algorithm and their biochemical functions.

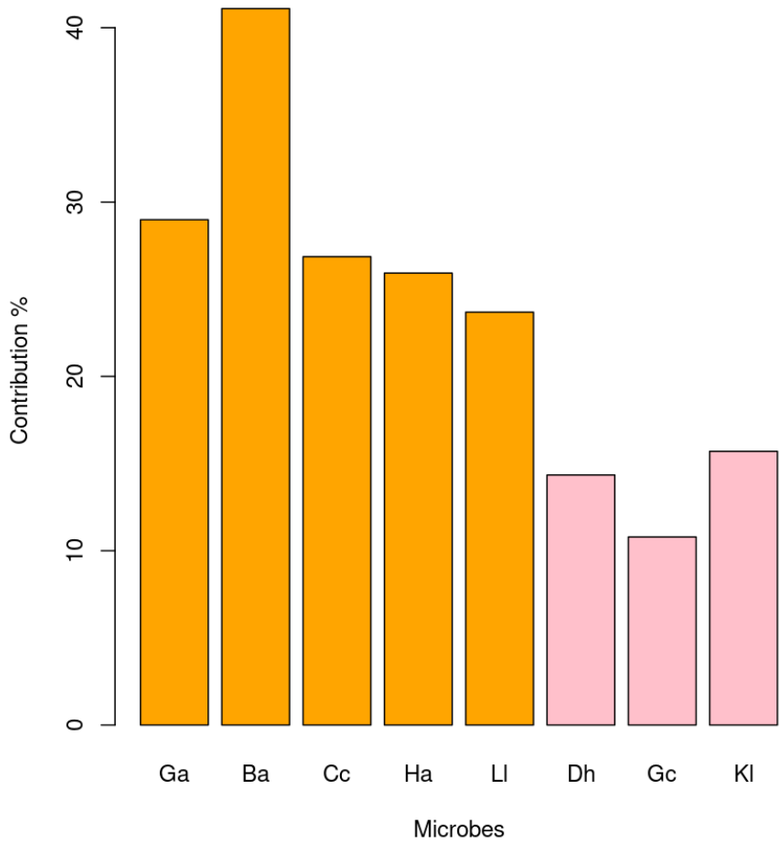
Microbes' contribution to the signatures

The comparison of the bacterial and yeast's expression profiles in both training and test transcriptomics data showed that yeasts express about two to three times more genes (mRNAs) than bacteria. It was expected that the yeasts would contribute more to

the models' signatures, however, the EN feature selection has retained more bacterial than yeast's genes as shown in figure 2.2. In FTS-selected gene set, the percentages of selected genes from each bacteria were: *Ga* 13 %, *Ba* 20 %, *Cc* 10 %, *Ha* 10%, *Ll* 11 %, *Se* 13 %, *Dh* 5 %, *Gc* 5 % and *Kl* 6 %. This holds true also in RTS where bacteria contributed more then yeasts to the selected gene set: *Ga* 30 %, *Ba* 41%, *Cc* 27 %, *Ha* 26 %, *Ll* 24 %, *Dh* 14 %, *Gc* 11 % and *Kl* 16 %. In both selection procedures, Ba species had the highest contribution. Note that the percentages dont' sum up to 100 because each microbe's percentage is independent from the others, i.e. the number of selected genes belonging to each microbe was divided by the total number of its expressed genes, and not by the total number of selected genes (see Methods).



(a)



(b)

Figure 2.2: Contribution percentage of each species to the selected gene sets from both FTS (a) and RTS (b). Orange bars: bacteria. Pink bars: yeasts.

This suggests that, in this specific case, i.e, experimental cheese, bacteria contribute to flavor formation more than yeasts. These observations are in line with the correlation results (Figure 2.3) where flavor amounts were more correlated with the growth of bacteria than yeasts. In other words, the increase or decrease of bacterial metabolic activity due to a decreases/increase of their number, induces a significant variation of the cheese flavor profile.

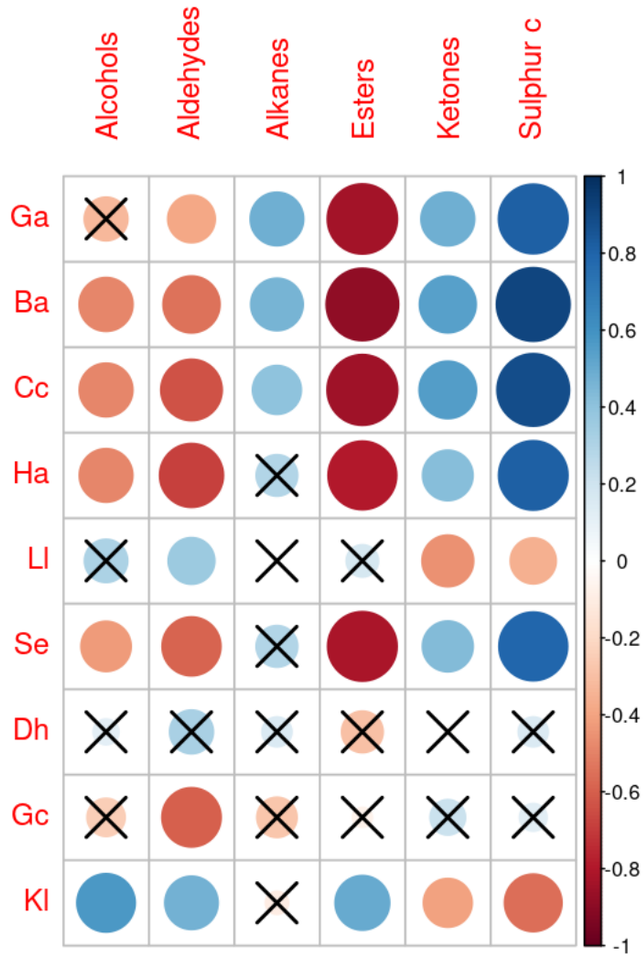


Figure 2.3: Spearman correlations between microbial growth and metabolites amount. Black crosses represent non-significant correlations. The areas of the circles and the colors' intensity are proportional to the absolute value of the correlation coefficients.

2.0.4 Conclusion

The training of predictive models on highly dimensional and imbalanced data is challenging. To cope with these problems, we performed three modeling procedures to build classifiers that can predict the flavor profile of cheese from its microbial expression profile. To rank our modeling approaches, we averaged the obtained performances, specifically, the balanced accuracy (BA) and the specificity, to get an overall estimation. The considered models were those which were trained on the RTS and validated on the independent test set, as this is the most robust way to assess a predictive model.

In terms of accuracy, the Random Forest classifiers trained on predictors selected by EN were the most robust (Average BA = 0.83), then, come the RF models trained on predictors selected by EN with an average BA = 0.77, and, lastly, the EN classifiers with BA = 0.47. In terms of their capability to correctly predict the minority class "High", the RF models trained without prior feature selection were the best (Average Specificity = 0.60), then come the RF models trained on selected features with an average specificity = 0.49. The EN models had the lowest average specificity which was equal to 0.30.

Statistically, the model performances are satisfying given the challenging data they were built on. In terms of biological relevance, we showed the EN feature selection and regularization method could select genes that are significantly associated to metabolic pathways.

2.0.5 Discussion

Here we review the work and try to point out some issues and propose alternatives to overcome them.

First, the data we started from were not very suited for training predictive models as they were sparse, i.e, having few data points (34 observations) and a large number of predictors (FTS: 4145 genes; RTS: 1062 genes), in addition to the existence of class imbalance. All these aspects induces many problems such as over-fitting, predictors multicollinearity and decreased specificity. Furthermore, as we couldn't train robust regression models, we opted for classification ones but no human expert labeling of the numerical values were available for these experiments. Therefore, to assign a label "High/Low" to each metabolite amount we used a pure statistical method, i.e. K-Means clustering, by assuming that that the metabolites amounts can be grouped into two groups, and this doesn't hold true for all the metabolites in the two experiments. For some of metabolites, it's hard to clearly distinguish two distributions (e.g. Esters in Figure 2.1 a). This approximate clustering of the outcome variable comes at a cost, it resulted in a class imbalance for most of the metabolites, which in turn makes the training and the validation of the models more challenging.

Nevertheless, the performances of the cross-validated models were very satisfying with an accuracy ranging from 68 % to 94 %. Furthermore, we provided also a more robust validation using a totally independent test set, overall, the performances of these models were satisfying and their predictions were significant (AUC > 0.5) It's true that most of them suffered from a low specificity, but in average, their accuracy can be qualified as good.

The usefulness of such models that are in general accurate but less powerful when it comes to predict the minority class, is a topic of discussion in the scientific community [176, 177]. Should we discard them or they are still useful? Actually, it will depend on the context and the reason they are used for. If one aims at correctly predicting the majority class then these models can be effective, but, if one is interested in the rare events (minority class) because of their importance, then low-specificity models will not be so useful.

One of the objectives of this study was to demonstrate that estimation of cheese flavor profile from the gene expression of the microbial communities is feasible. These straightforward estimations can make cheese quality control faster and cheaper as compared to the traditional metabolomics and sensory approaches. To this aim, more robust and generic models can be built if cheese big data will be made available, especially, transcriptomics and metabolomics data. Therefore, there is a need to create such databases where omics-data of different cheeses are merged and made available for different analyses, especially, machine learning. These models can be then used by industrials during cheese making processes for a more effective quality control.

Elastic Net regularization was successfully used to select genes associated to phenotypic traits. Torang et al. [178] used EN to find gene signatures that can distinguish between the types of immune cells. In our study, we could select genes significantly related to metabolic pathways. We can make use of such algorithms in the field of metabolic engineering. In a nutshell, metabolic engineering (ME) aims at modifying the cells regulatory and metabolic pathways to obtain desirable functions and products [179]. This technique has been used to create modified microorganism used for bioremediation (degradation of xenobiotics) and production pharmaceutical. In food industry, modified microorganisms have been used for many tasks, such as improvement of starch utilization in bakery and improvement of lactic acid production in dairy products [180, 181]. ME relies on the identification of key genes related to the traits and functions one wants to improve, therefore, machine learning algorithms can be effective tools to identify these genes-traits relationships.

2.0.6 Supplementary data

Alcohols		Aldehydes		Alkanes		Ketones		Esters		Sulphur compounds	
sum of peak area/ug	class	sum of peak area/ug	class	sum of peak area/ug	class	sum of peak area/ug	class	sum of peak area/ug	class	sum of peak area/ug	class
10124523145	High	376378343.5	Low	37064689.5	High	378081950.5	Low	158706511.5	High	0	Low
11722959640	High	413883114.5	Low	18223824	Low	492386352	Low	340132217	High	0	Low
6615632990	Low	310272482.5	Low	16769897.5	Low	445646454.5	Low	102917252	Low	0	Low
3074211745	Low	84797113	Low	13587024.5	Low	569999894.5	Low	67248738	Low	123547564.5	Low
2797998855	Low	104350683.5	Low	16994756	Low	561304797.5	Low	37682911.5	Low	109888628	Low
3085057341	Low	98055634	Low	16499981.5	Low	501061474	Low	61164502	Low	201272543	Low
4033984593	Low	86006394	Low	4133625	Low	1226839020	High	37385971	Low	726802143	Low
3361234931	Low	74004569.5	Low	3059935.5	Low	1328949659	High	65107609.5	Low	794680513.5	Low
4024448307	Low	86236954	Low	4438497.5	Low	1376454111	High	62308150	Low	752103830	Low
2515145265	Low	74868468	Low	3478566	Low	2309511281	High	10482687	Low	3227065375	High
2814140358	Low	53633946	Low	2860838.5	Low	2305940112	High	6824870	Low	3139001154	High
2781125030	Low	51100156	Low	3487098	Low	2049771036	High	10772413.5	Low	3066026893	High
9225949284	High	1483503114	High	30400891.5	High	162960272.5	Low	10989275	Low	955791	Low
12606926370	High	1788154904	High	45544255.5	High	192401565	Low	93795479	Low	1002682.5	Low
13516201735	High	1887035141	High	34057368	High	192478713	Low	27157158.5	Low	814631.5	Low
6690630861	Low	486731100	Low	17670482.5	Low	452212088	Low	3992940	Low	1593188.5	Low
5823500255	Low	351003232.5	Low	14088737.5	Low	481889280	Low	1536631	Low	857833	Low
9022873089	High	472918795.5	Low	18706955	Low	771533164.5	Low	2300312.5	Low	1295060.5	Low
2352617804	Low	125162729	Low	34340061	High	435213228	Low	0	Low	48822905	Low
2383854839	Low	140725146	Low	50563515.5	High	1405866200	High	0	Low	78696013.5	Low
3272912283	Low	226334561	Low	51422275.5	High	688505413.5	Low	0	Low	51717399.5	Low
2296836653	Low	163956101	Low	71030591.5	High	1158417474	High	0	Low	368726016.5	Low
2282536194	Low	190784532.5	Low	47587467	High	955469147	Low	0	Low	448076526	Low
2063983500	Low	113892826.5	Low	44120082.5	High	894460334.5	Low	0	Low	425189852.5	Low

Table 2.14: Labeling of the numerical metabolite amounts in the test set.

Chapter 3

Network diffusion analysis to elucidate antimicrobial resistance mechanisms of *E. coli* and reveal potential drug targets

3.0.1 Introduction

The continuous increase of antimicrobial resistance (AMR) is of concern for public health. Out of 4.95 millions deaths caused by microbial infections, it was estimated that bacterial AMR was directly responsible for 1.27 million global deaths in 2019 [182]. AMR is defined as the ability of microbes to survive and bypass the effect of different antimicrobial agents such as antibiotics, disinfectants, and food preservatives. The misuse and overuse of these agents can cause bacteria, viruses, fungi and parasites to adapt by developing protective mechanisms such as mutations, metabolic adaptations and secretion of anti-antimicrobial molecules. The microbes can become totally resistant or less susceptible against the antimicrobial agents, as a result, the disease spread increases and leads to severe illness, disability and deaths [183, 184, 185].

The emergence of multidrug-resistant or pan drug-resistant Gram-negative bacteria suggests the need to focus the research of both the scientific community and pharmaceutical companies on the development of new antimicrobials. In the past, the approach to drug discovery was essentially based on the analysis of one or two levels of biological information linked explicitly to the target and mechanism of action of the molecule within a biological model. More recently, with the progress in omics and computational sciences, the scientific community has reached an increasing awareness that genes and proteins are not acting as standalone molecules, but they interact on multiple hierarchical levels as complex networks [186, 187, 188]. In the cell, proteins are organized in complex structures and collaborate to perform biological functions. A comprehensive mapping of interactions between genes in the genome is a relevant information to understand such processes . Protein–protein interaction (PPI) networks aim to grasp this complex pattern of interactions by modeling individual proteins as vertices, and their relationships as undirected edges. In general, coding genes are considered to have a one-to-one relationship with proteins. PPI information can be retrieved from a variety of resources based on known and/or computationally predicted interactions. In on-line resources such as STRING [189], GeneMANIA [190], FunCoup [191] and ConsensusPathDB [192], experimental data are integrated with interaction prediction algorithms, thus aiming for high comprehensiveness and coverage.

Meaningful biological insights can be extracted from biological networks through different network-based analyses. Network diffusion (ND) is used in many contexts such as gene prioritization, function prediction, survival prediction and disease sub-typing [193]. ND relies on the propagation of information (signal) from already characterized source nodes (”seed” genes, proteins, metabolites...) through the network. Basing on the guilt-by-association principle [194], nodes that are more adjacent to the source nodes will be prioritized and will accumulate more signal after convergence, and thus, they will be more likely related to the phenotypic trait or the function of interest [195].

In biological studies, ND has been used for different purposes such as the prediction of functional associations of unannotated gene sets [196], the prediction of protein functions [197], and the identification of cancer gene mutations [198]. Recently, ND was applied to immune-related protein interaction networks to identify key proteins that can be new COVID-19-drug targets [199].

Objective of the work

In this work, we aimed at deepening the knowledge of AMR mechanisms in *Escherichia coli* in order to provide relevant biological insights that can contribute to the development of new antimicrobial therapeutic strategies.

Summary

We propose a systems biology approach to identify genes and biological pathways associated with AMR, by mapping known AMR-related genes from CARD and PointFinder databases into the *E. coli* protein interactome. Through a network diffusion algorithm already applied in similar contexts, we identified network modules, consisting in a list of genes and pathways, in part already known to be involved in AMR mechanisms and in part new, out of which we selected relevant gene candidates for *in vitro* susceptibility testing knockout mutants against nine different antibiotics. Compared to the wild-type *E. coli BW25113*, the mutants $\Delta uhpB$, $\Delta mdaB$, $\Delta rpmG$ and $\Delta rplA$ showed a significant shift in their anti-microbial susceptibility to streptomycin, ciprofloxacin, ampicillin, tetracycline and chloramphenicol. Both experimental (susceptibility tests) and statistical (ORA) validations demonstrated the effectiveness of network-based *in silico* approaches such as ND in discovering relevant genes associated to AMR in *E. coli*.

Our results contribute to a better understanding and characterization of antimicrobial resistance in *E coli*, furthermore, the *in vitro* validated genes represent new putative drug targets.

3.0.2 Methods and Materials

Protein-Protein Interaction Networks (PPI)

For tour analysis, we used the PPI network of the reference organism *E. coli K12 MG1655* available on STRING-v11.5 [200]. The PPI undirected interactions were collected based on the "b number" *E. coli* gene identifiers for each node. A total of 4053 nodes and 33656 edges were retained, considering only relationships between proteins characterized by a high combined confidence score (score ≥ 0.7) as provided by STRING website.

Mapping known antimicrobial resistance genes on *E. coli K-12 MG1655* PPI

CARD (v3.2.7, https://card.mcmaster.ca) [201] and PointFinder (v.4.1.0) [202] databases were selected as comprehensive databases of AMR-related genes. In total, we collected a list of 34 AMR-related genes from the two databases that could be mapped onto the *E. coli's* PPI (Table 3.1): 32 genes from CARD, 2 from PointFinder. These genes were used as "seed genes" for the network diffusion procedure.

b#	Gene Name	CARD	PointFinder
B0463	AcrA	✓	
B0464	AcrR	✓	
B0543	EmrE	✓	
B0578	NfsB	✓	
B0842	MdfA	✓	
B0851	NfsA	✓	
B0929	OmpF	✓	
B1093	FabG	✓	
B1288	FabI	✓	
B1530	MarR	✓	
B1782	GyrA	✓	✓
B2231	GlpT	✓	
B2240	PtsI	✓	
B2416	ParC	✓	✓
B3019	ParE	✓	✓
B3030	FolP	✓	✓
B3177	MurA	✓	
B3189	UhpT	✓	
B3669	UhpA	✓	
B3699	GyrB	✓	✓
B3806	CyaA	✓	
B3912	CpxR	✓	
B3987	RpoB	✓	
B4036	LamB	✓	✓
B4062	SoxS	✓	
B4063	SoxR	✓	
B4150	AmpC	✓	✓
B4396	Rob	✓	
B4113	BasR/pmrA		✓
B4112	BasS/pmrB		✓
B0084	FtsI(PBP3)	✓	
B3339	TufA/EFTu	✓	
B3980	TufB/EFTu	✓	

Table 3.1: List of CARD and PointFinder genes mapped to *E. coli K-12 MG1655* protein-protein network. The column "Gene_name" reports the genes' SYMBOL IDs and their synonyms, if any.

The seed genes list comprises the genes annotated with the prefix "Ecol_" according to CARD annotation. Genes related to AMR in *E coli* without this prefix are left as a validation gene set to assess how well the diffusion approach can recover genes that are already well known to be involved in AMR.

Network diffusion Analysis

To identify genes and pathways associated with AMR, we employed the Bersanelli et al. [51] network diffusion algorithm from the R package *diffuStats* [203]. This algorithm consists in a random walk with restart, in which the initial 34 AMR gene list and the *E. coli K-12 MG1655* gene interaction network represent the inputs.

The network diffusion process simulates fluid dispersion in the network, in which the seed genes act as fluid (information) sources. The diffusion starts with an initial vector S_0 of scores for each node: $s_0(seed\_gene) = 1$, $s_0(not\_seed\_gene) = 0$. Then, given the network's adjacency matrix A, the algorithm iteratively propagates the information to the *non_seed_genes* to produce a score vector S^* at convergence for all the nodes:

$$s_{t+1} = \alpha A \cdot s_t + (1 - \alpha)s_0$$

where α defines the probability for the fluid to be retained by the source nodes.

It controls how much information is kept in the nodes versus how much tends to be spread through the network. The convergence is reached when there is no significant difference between s_{t+1} and s_t:

$$|s_{t+1} - s_t| < 10^{-6}$$

In our study, we assign an initial score of 1 to the seed genes only. Due to the diffusion process, the genes with high connectivity degree, i.e. the network hubs, may accumulate more fluid, thus acquiring a high diffusion score s^* at convergence, only because of their central position in the network. Thus, to mitigate this hub effect we consider the smoothed version of the s^* as previously described [51].

To select the top-ranking genes, sorted by their diffusion score s^* after convergence, a permutation procedure was applied by running 1000 diffusions with 34 randomly assigned seed genes at each iteration. A p-value for each true s^* score was then computed (p_val = proportion of random scores $\geq$ true s^* score). Genes having an s^* score with p-value ≤ 0.01 were selected as top-ranking AMR-related genes. In total, there were 127 selected genes (including the 34 seed genes) with a significant diffusive score.

Community detection: Louvain algorithm

From the initial *E. coli* PPI, we extracted a sub-network corresponding to the top-127 genes selected by the network diffusion step. We run the Louvain algorithm [65] implemented in *igraph* R package [204] to detect communities within these sub-network and try to associate them with specific biological pathways in which these genes are involved. We chose this method due to its proven good performances in detecting communities in complex networks [60].

The Louvain method is so called because all the authors were connected to the Catholic University of Louvain in Belgium [205]. It is an unsupervised community detection algorithm based on modularity optimization, i.e, it creates clusters such that the forming nodes have the optimal modularity. The algorithm has two phases:

- **Phase 1**: all the nodes of the network are considered as starting communities, then, local communities are built by linking each individual node with its neighbors. Secondly, each of the initial nodes are moved from their local community to a host-community, i.e, another local community to which the moved node doesn't belong. If the moved node increases the modularity of the host community then it's kept there. This operation is done for all nodes until obtaining local communities yielding an optimal modularity score. Finally, these communities are aggregated to build super-nodes.

- **Phase 2**: starting from the super-nodes a new network is built. The super-nodes are interlinked to each other if there exists at least one edge connecting two nodes from two different super-nodes.

  Now that a new network is obtained, the **Phase 1** procedure is repeated.

- **Convergence of the algorithm: Phase 1** and **Phase 2** are repeated iteratively until no improvement of modularity is observed. The resulting communities will be then the output of the algorithm.

Pathway enrichment analysis

KEGG pathway annotation [149] was used for the identification of over-represented pathways. The selected genes were used to perform an over-representation analysis (ORA) based on hypergeometric distribution via the R package *clusterProfiler* [206]. The p-values of the enrichment analysis were corrected for multiple testing by using the Benjamini-Hochberg post-hoc method [207]. An FDR ≤ 0.05 value was chosen to define the significantly over-represented pathways.

Validation of identified genes

Out of the genes prioritized by the network diffusion algorithm, 13 genes were selected for *in vitro* evaluation as potential antimicrobial resistance targets. The knockout strains used in this study were obtained from the KEIO collection [208], including the parent strain *Escherichia coli BW25113*. All strains were initially streaked on LB agar (Sigma Aldrich), prepared according to the manufacturer's instructions, and incubated at 37°C. Isolated colonies from these plates were then cultured in LB broth (Sigma Aldrich) for subsequent antibiotic susceptibility testing.

Antibiotics Susceptibility Test

The assessment of antibiotic susceptibility in *E. coli* isolates was conducted utilizing the Kirby-Bauer disc diffusion method following the protocols outlined by the Clinical and Laboratory Standards Institute [209]. A total of nine antimicrobial agents were assessed: Ampicillin (10 μg), Chloramphenicol (30 μg), Ciprofloxacin (5 μg), Fosfomycin (200 μg), Penicillin G (10U), Polymyxin B (300U), Spectinomycin (10 μg), Streptomycin (10 μg), and Tetracycline (30 μg). Initially, the *E. coli* isolates were cultured in nutrient broth and then incubated at a temperature of 35 $\pm$ 2 °C for 18-24 hours. The bacterial suspension was subsequently standardized to a 0.5 McFarland turbidity standard, resulting in a concentration of around 108 CFU/ml. Employing cotton swabs, the bacterial suspension was uniformly distributed on Mueller-Hinton agar plates and allowed to air dry for 15 minutes. The antibiotic discs were positioned on the agar surface with a minimum distance of 30 mm between each disc. Subsequently, the plates were inverted and aerobically incubated at a temperature of 35 $\pm$ 2 °C for 16-18 hours. The zones of inhibition were quantified using an automated colony counter (Interscience Scan500) and interpreted in line with the recommendations provided by the CLSI [210]. To ensure quality control, the *E. coli ATCC 25922* strain was utilized. The bacteriological media were procured from HiMedia Laboratories in Mumbai, India, while the antibiotic discs were sourced from Thermo Scientific™ Oxoid™.

For each tested antibiotics, the average inhibition diameters of the mutants were compared to the wild type using an unpaired Student' s T test. The p values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg post-hoc correction for multiple tests [207]. The statistical analysis was carried out using R [154].

3.0.3 Results

Network-based and pathway analyses

First, the 34 AMR-related seed genes extracted from CARD and Pointfinder databases (Figure 3.1) shows the inter-connectedness of these genes. The over-representation analysis (ORA) of these seed genes resulted in six inter-connected KEGG pathways: beta-Lactam resistance, Cationic antimicrobial peptide (CAMP) resistance, Two-component system, Fatty acid biosynthesis and metabolism and biotin metabolism. These results reflect the complexity of the antimicrobial mechanisms which can involve more than one biological pathway [211, 212].

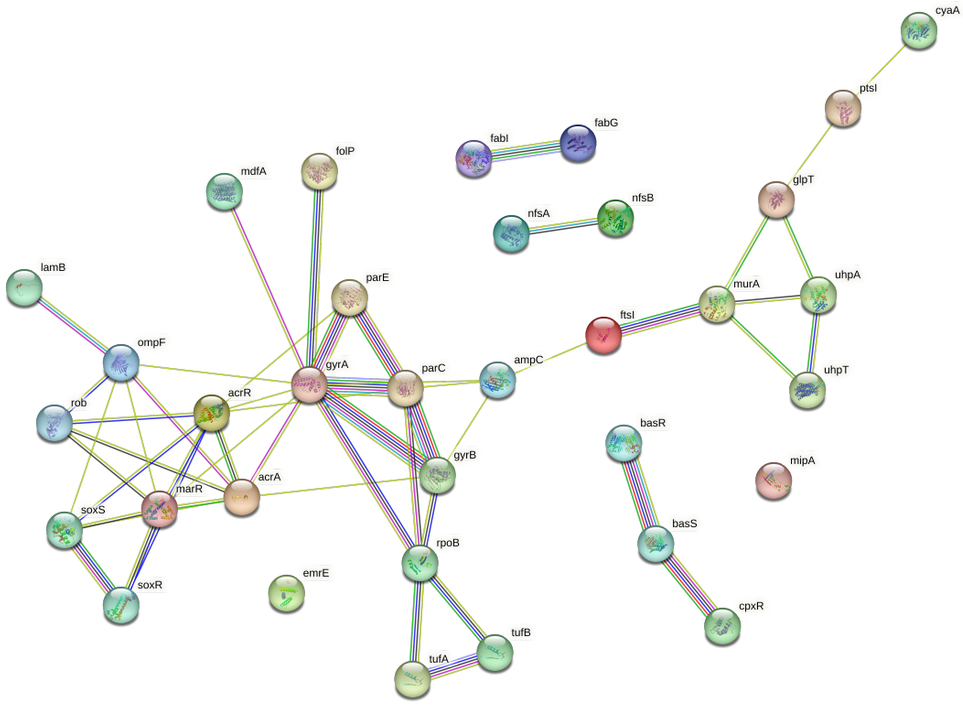
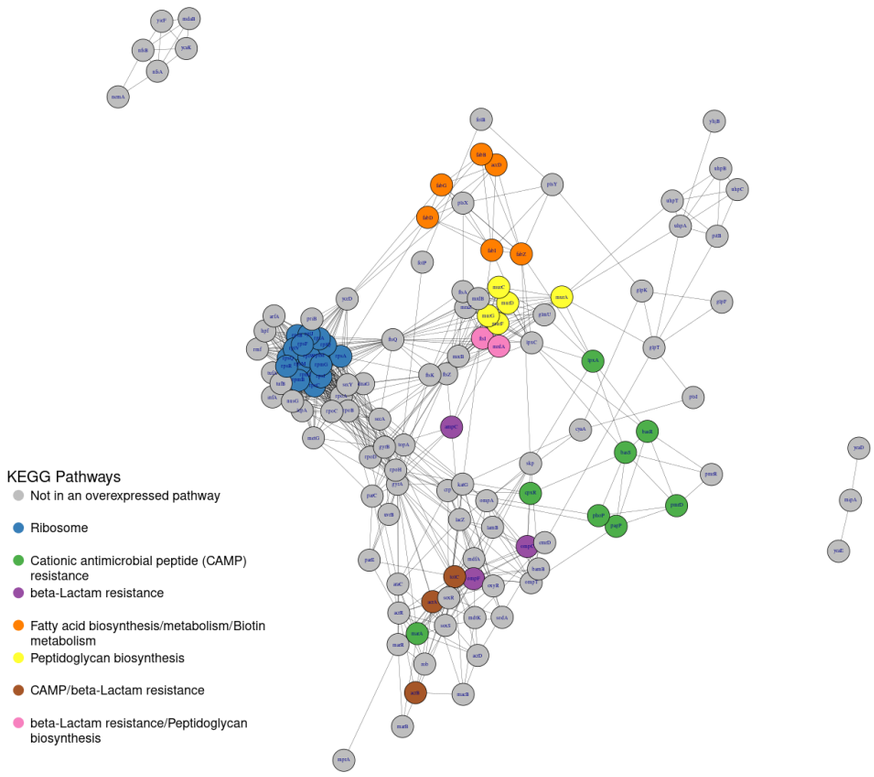


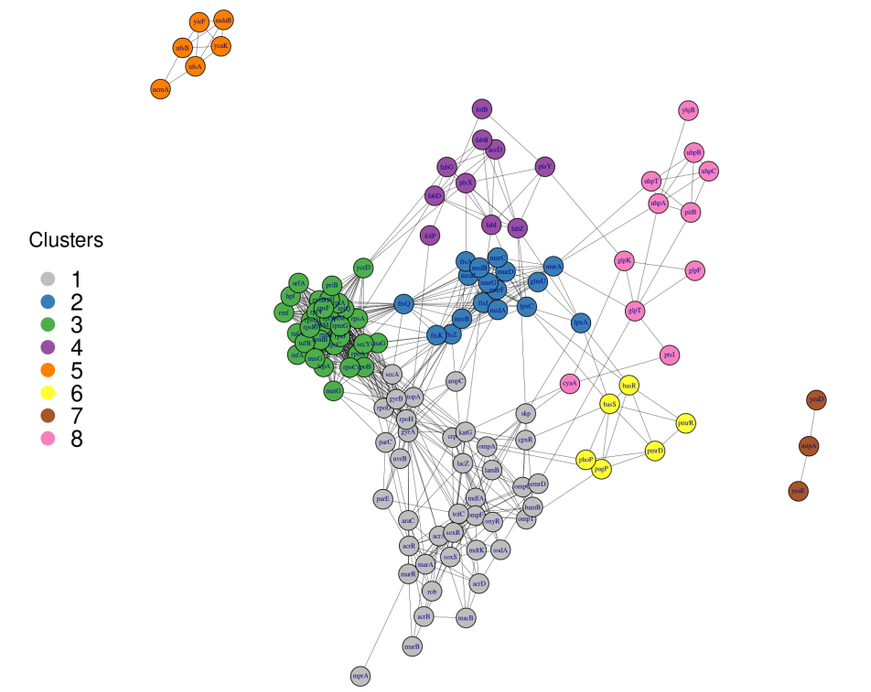
Figure 3.1: Subnetwork of the 34 known antimicrobial genes (seed genes) with the links found in *E coli* PPI STRING interactome.

The ND generated a list of 127 genes associated to AMR: 34 are the original seed genes, while the remaining 93 constitute the novel achievement of our approach. We considered the PPI sub-network induced by the 127 gene list, consisting of one big, connected component of 117 nodes and two small components of 6 and 3 nodes, and one isolated node. ORA applied to these 127 genes resulted in seven significantly enriched pathways, namely, Ribosome, CAMP, beta-lactam resistance, Fatty acid biosynthesis, Peptidoglycan biosynthesis, Fatty acid metabolism and Biotin metabolism (Table 3.2). These results overlap with the ones obtained by using ORA applied to the 34 seed genes only. This suggests that the 93 prioritized genes added by the diffusion algorithm were consistent with antimicrobial resistance pathways in *E. coli*.

Mapping these pathways to the eight network clusters defined by the Louvain community detection algorithm (Figure 3.2 a and b) resulted in a partial overlap: Ribosome pathway overlaps with cluster 3, CAMP with cluster 6, beta-lactam resistance with cluster 1, Fatty acid biosynthesis/metabolism and biotin metabolism, which share their genes, overlap with cluster 4, Peptidoglycan biosynthesis overlap with cluster 2. These cluster are partially consistent with biological pathways due to the overlapping nature of these latter. For instance, CAMP, peptidoglycan and beta-lactam resistance as well as fatty acid metabolism/biosynthesis and biotin metabolism are pathways that have several genes in common: genes and proteins can have multiple functions, and thus, be involved in multiple pathways.



(a)



(b)

Figure 3.2: (a) The Induced sub-network of *E. coli k-12 MG1655* protein-protein network with 127 genes colored by membership to one of the 8 clusters identified by the Louvain community detection algorithm. (b) Plot of the same network with genes colored by KEGG pathway annotation. The isolated gene *emrE* is not shown.

Before reviewing the identified genes individually (next section), we analyzed their relation to AMR in *E coli* at the pathway level by considering the over-represented pathways (Table 3.2). Ribosome was the most significant and largest pathway with seventeen 30S and 50S ribosomal subunit proteins responsible for decoding mRNAs and control of translation fidelity, and the catalysis of protein synthesis, respectively [213]. Interestingly, none of them was a seed gene. We investigated their first-order neighborhood and we found four seed genes: *tufA*, *tufB*, *rpoB* and *gyrA* (Figure 3.3). These links have allowed the information flow to reach ribosomal genes and made them appear in the top-ranking list. The diffusion process was thus able to uncover the relation between AMR and ribosomes which represent important targets of many antimicrobial therapeutical strategies, such as miscoding using streptomycin and paromomycin [214], minimization of ribosomal mobility [215], blockage of the protein exit tunnel [216].

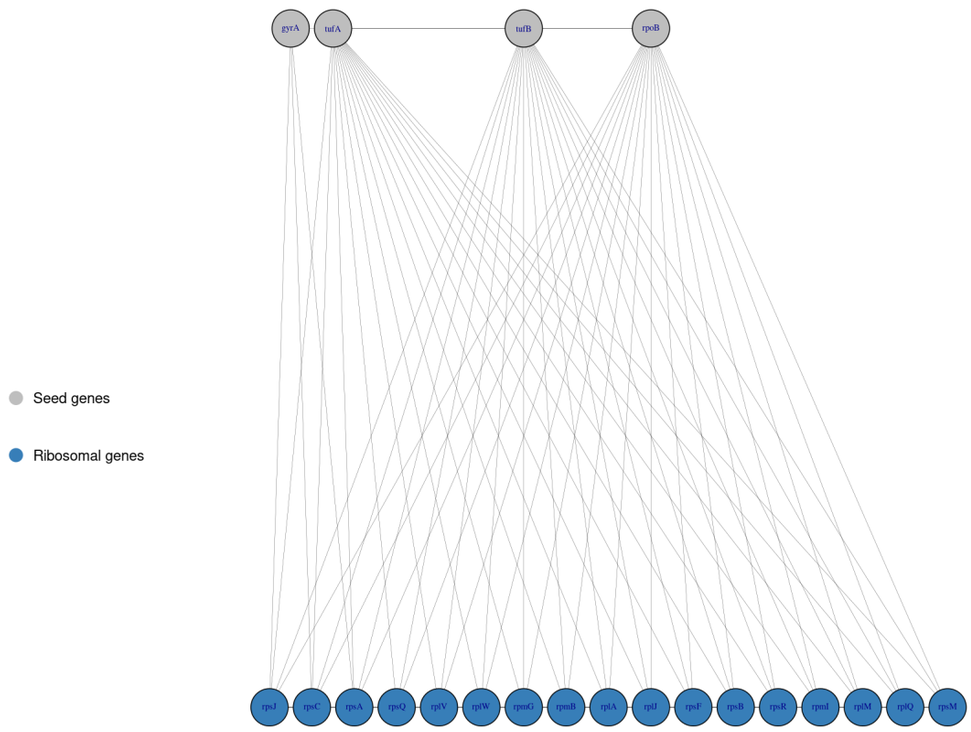


Figure 3.3: A sub-network showing the links between the seed genes *tufA*, *tufB*, *rpoB* and *gyrA* with the ribosomal genes prioritized by the network diffusion algorithm.

CAMP resistance pathway was the second largest enriched pathway (11 genes). The cationic peptides (CPs) are antimicrobial components naturally expressed by animals, plants and even bacteria [217]. Through electrostatic interactions, they bind the outer layer of the bacterial cytoplasmic membrane and induce the lysis of the targeted microbial cell [218]. Bacteria can acquire resistance against CPs through different mechanisms such as the modification of the cell surface structure and its net charge in gram-negative bacteria to alter CPs binding [217]. Another CAMP resistance mechanism observed in *E. coli* and *S. aureus* relies on the trapping and proteolytic degradation of CPs by producing of metalloproteinases [219]. Moreover, it has been shown that *E coli* and other bacteria have developed another strategy to survive in CP-rich environments. It consists in the release of negatively charged capsular polysaccharides to neutralize and titrate CPs by means of electrostatic interactions [220].

Beta-lactams (BL) are antibacterial designed to disrupt the peptidoglycan structuration to provoke cell wall defects that lead to the lysis of the targeted bacteria.

Beta-lactam resistance (3rd over-represented pathway) is one of the early characterized and most successful AMR strategies used by bacteria [221], which acquire resistance against BL, such as penicillins, through different mechanisms: (1) Inactivation and destruction of the antimicrobial through beta-lactamases (2) Decreased penetration of the antimicrobial to the target site or alteration of the target site itself (3) Excretion of BL molecules through efflux pumps [222].

Finally, the over-represented pathways fatty acid, peptidoglycan and biotin biosynthesis/metabolism can contribute to AMR resistance through global cell adaptations of the bacteria such as metabolic adaptations and cell envelop homeostasis. Fatty acids biosynthesis is a key pathway in bacterial cell growth; therefore, it represents an important antimicrobial target [223]. *E. coli* can acquire resistance to antibacterial agents that inhibit its lipid synthesis such as triclosan by altering its target, the enoyl-[acyl-carrier-protein] reductase *fabI*. A missense mutation in *fabI* hinders triclosan activity by reducing the binding affinity of the complex *fabI*-triclosan [224]. Survival strategies relying on membrane homeostasis are found in many pathogens to increase their fitness in the presence of environmental stressors such as antibiotics [225]. It has been observed that increased tPMP (thrombin-induced platelet microbicidal protein) resistance in *S. aureus* is due to a higher cell membrane fluidity caused by a preponderance of longer chain, unsaturated fatty acids [226].

Peptidoglycan sacculus (PS) is an elastic and net-like polymer that surrounds the cytoplasmic membrane in most bacteria. It contributes to the preservation of the cell integrity during growth and division and the protection of the bacteria against environmental challenges such as osmotic stress [227]. PS structuration is disrupted by BL that target the penicillin-binding proteins responsible for the synthesis of the 4→3 peptidoglycan cross-linking. To cope with BL effects, *E. coli* expresses the L,D-transpeptidase *YcbB* which catalyses an unusual 3→3 peptidoglycan cross-linking to maintain cell wall integrity [228].

Pathway	**FDR**	**# Genes in Network**	**Gene IDs**	**# Total Genes in Pathway**	**Seed Genes**
Ribosome	4.43e-07	17	rpsB, rpsA, rpmI, rplM, rplQ, rpsM, rpsQ, rpsC, rplV, rplW, rpsJ, rpmG, rpmB, rplA, rplJ, rpsF, rpsR	78	-
CAMP resistance	2.3e-06	11	lpxA, acrB, acrA, pagP, phoP, marA, pmrD, tolC, cpxR, basS, basR	36	AcrA, cpxR, basS, basR
Beta-lactam resistance	2.3e-06	8	ftsI, acrB, acrA, mrdA, ompF, ompC, tolC, ampC	17	ftsI, acrA, ompF, ampC
Fatty acid biosynthesis	8e-05	6	fabZ, fabD, fabG, fabI, accD, fabB	13	fabG, fabI
Peptidoglycan biosynthesis	3.7e-04	7	ftsI, murF, murD, murG, murC, mrdA, murA	24	FtsI, murA
Fatty acid metabolism	1.3e-03	6	fabZ, fabD, fabG, fabI, accD, fabB	21	fabG, fabI
Biotin metabolism	1.4e-02	4	fabZ, fabG, fabI, fabB	14	fabG, fabI

Table 3.2: Results of the pathway enrichment analysis (ORA) applied to the 127 genes including 34 seed genes and 93 AMR-associated genes newly identified by the network diffusion algorithm.

Literature validation of the prioritized AMR genes

In table 3.3 we listed the nodes (genes) with the highest diffusive score s^* in each over-represented pathway (excluding original seed genes). The complete list of the identified genes and their annotation is provided in the supplementary table 3.5. Except rpmB, all genes in table 3.3 are already known to be associated to AMR in many microorganisms including *E coli*. For each gene, we reported the information source (literature and/or CARD) where a detailed description of the AMR mechanism is provided.

Gene	s^*	**Annotated Pathways**	**Source**
uhpB	0.26	No	[229]
uhpC	0.25	No	[229]
rpmB	0.027	Ribosome	
rpsj	0.026	Ribosome	[230], CARD
rpmG	0.026	Ribosome	[231]
marA	0.1	CAMP resistance	[232], CARD
pmrD	0.05	CAMP resistance	[233, 234]
ompC	0.04	Beta-lactam resistance	[235, 236], CARD
fabD	0.05	Fatty acid biosynthesis/metabolism/Biotin	[237]
rabB	0.04	Fatty acid biosynthesis/metabolism/Biotin	[238]
fabZ	0.04	Fatty acid biosynthesis/metabolism/Biotin	[239]
murC	0.04	Peptidoglycan biosynthesis	[240]
tolC	0.09	CAMP resistance/Beta-lactam resistance	CARD
acrB	0.07	CAMP resistance/Beta-lactam resistance	CARD
mrdA	0.04	Beta-lactam resistance/Peptidoglycan biosynthesis	[241]

Table 3.3: Most relevant genes considering the highest values of diffusion score Sx (excluding seed genes) and their biological pathways.

uhpB and *uhpC* are inner membrane proteins belonging to the phosphorelay system *uhpB-uhpC-uhpA* responsible for sensing glucose-6-phosphate and its accumulation into the bacterial cells. Chromosomal mutations of *uhpB* and *uhpC* confer resistance to fosfomycin in *E. coli CFT073* [229].

30S and 50S ribosomal subunit proteins constitute the largest over-expressed pathway (Ribosome, 17 genes, Table 3.2). Under high tigecycline concentrations, nine populations of *E. coli BW25113* were able to grow due to a reduced susceptibility conferred by a mutation in the ribosomal S10 protein rpsJ [230].

marA is transcriptional dual regulator which is part of the multiple antibiotic resistance (MAR) chromosomal locus in *E coli*. It has a positive regulatory activity on *acrAB–tolC* efflux pump system that confers a multi-drug resistance by an active transport of antimicrobial molecules outside the cell [242, 243]. In a recent study [232], two highly resistant *E. coli* strains EV18 and EVC (resistant to NF and chloramphenicol, respectively) have been found to have a low cytoplasmic pH compared to the sensitive wild-type. It has been shown that MAR operon can contribute to the decrease of cytoplasmic pH, which is considered as a basal microbial protection mechanism against antimicrobials.

pmrD is a signal transduction protein, when over-expressed in *Salmonella Typhimurium* it confers resistance to polymyxin B [233]. *E. coli* has a pmrD homologue which is required for the modification of the lipopolysaccharide structure of lipid A. This modification has been shown to promote resistance against CPs [234].

ompC is one of the major outer membranes porins in *E. coli* playing an important role in protecting the bacterial cell against harmful physical and chemical stressor such as toxins and antibiotics [244]. The over-expression of membrane porins in *E. coli* leads to reduced susceptibility to BLs and other antibiotics [235, 236]. BL antibiotics such as carbapenem and diazabicyclooctane are effective inhibitors of the penicillin-binding proteins (PBPs) involved in bacterial cell wall biosynthesis. Several *E. coli* clinical isolates are less susceptible to these antibiotics because of the presence of mutations in the mrdA gene (PBP2) [241].

fabD, *fabB*and *fabZ* genes play essential roles in fatty acid biosynthesis and elongation in *E. coli* [245, 246]. *fadDB* mutants have been shown to be resistant against the calmodulin antagonist trifluoroperazine. This resistance could be due to a modification of the cell membrane permeability through a modulation of its fatty acid composition [237]. Thiolactomycin antibiotic (TLM) is a type II fatty acid synthesis inhibitor. In TLM-resistant *E coli* strain ANS1, a missense mutation of *fabB* gene (T1168G) leads to a functional *fabB* protein carrying a valine amino acid at position 390 (F390V). The valine's side chain hinders TLM effect by preventing the formation of TLM-fabBG complex [238]. Finally, chromosomal mutations in the dehydratase *fabZ* have been associated to a 200-fold increased resistance against LpxC inhibitors [239].

Because of their essential role in maintaining microbial cell integrity, genes involved in PS biosynthesis are interesting antibiotic targets, such as *murC*, an UDP-N-acetylmuramate-alanine ligase responsible for the addition of the first amino acid of the peptide moiety in the assembly of the monomer unit of peptidoglycan [247]. This gene has been targeted with a pyrazolopyrimidine antibiotic in both*E. coli* and *Pseudomonas aeruginosa*, and when the intracellular concentration of this antibiotic is optimal, *murC* enzymatic activity is inhibited in *E. coli* [240].

In vitro antibiotic susceptibility of mutants

In order to phenotypically investigate their role in AMR, 13 newly identified genes were selected for knockout experiments, based on the following criteria: 1) genes coding for proteins with functions potentially related to mechanisms of action of known antimicrobial agents; 2) Not seed genes 3) not essential genes (i.e. genes for which the knockout is not fatal for the bacterial cell). These genes were: *uhpB* (JW3643-KC), *mdaB* (JW2996-KC), *yieF* (JW3691-KC), *pitB* (JW2955-KC), *rplA* (JW3947-KC), *uvrB* (JW0762-KC), *rpmG* (JW3611-KC), *rpsF* (JW4158-KC), *nemA* (JW1642-KC), *ompC* (JW2203-KC) *ompT* (JW0554-KC), *yeaD* (JW1769-KC) and *yeaE* (JW1770-KC). *uhpB* is a sensor HK protein which controls production of the sugar phosphate transporter *uhpT* [248]. G469R mutation in the *uhpB* gene was associated to fosfomycin resistance [229]. The susceptibility of the knockout mutant to other antimicrobials with similar mechanisms of action than fosfomycin was not yet investigated. *mdaB* is a NADPH oxidoreductase that protects cells against quinonoid compounds [249]. It has been reported that the protein is able to confer resistance to the antibiotics DMP 840, adriamycin and etoposide [250]. *yieF* is a chromium reductase involved in bacterial tolerance to this heavy metal [251]. *pitB* is a phosphate transporter [252]. In the cell, ortophosphates were suggested to link heavy metals. Metal phosphates are transported out of the cell by *pitB* contributing to heavy metal resistance, [252]. *rplA* is a ribosomal protein. The knockout of the corresponding gene has been associated to zinc resistance [253]. Co-resistance against antimicrobials and heavy metals was described as synergistic with the potential to antimicrobial resistance [254]. In particular, heavy metals promote the spread of antimicrobial resistance genes and bacteria in the environment [255]. *uvrB* is involved in the SOS response associated to DNA biosynthesis and repair [256]. Another protein involved in DNA repair is *rpmG* which was associated to resistance against mitomycin C, a natural antimicrobial synthesized by *Streptomices caespitosis* and associated to DNA damage [231]. The *rpsF* gene codes for S6 ribosomal protein [257]. Mutations in ribosomal proteins have been already described as associated to erytromycin, spectinomycin and streptomycin resistance in *E. coli* [258]. The *yeaD* gene encodes for the d-exose-6-phosphate epimerase-like protein which is involved in galactose metabolism [259, 260]. Bacterial epimerases are involved in complex carbohydrates polymer that are used in cell wall and cell membrane [260]. Finally, the *nemA* gene encodes N-ethylmaleimide reductase in *Escherichia coli* [261]. The presence of the gene was associated to higher resistance of *E. coli* to acid hydrolysate of sugarcane bagasse [262]. The correlation between acid tolerance and antimicrobial resistance has been previously described [263]. *ompT* is a protease located on the outer membrane and participating to the adhesion of *E. coli O157:H7* to human epithelial [264].

Mutant	Streptomycin	Ciprofloxacin	Ampicillin	Tetracycline	Chloramphenicol
WT	I	I	S	S	S
uhpB	I (0.016)	I (0.012)	I*	S (0.016)	S*
mdaB	I (0.05)	I*	R*	S (0.05)	S (0.03)
rpmG	I (0.02)	S*	S*	S*	S (0.015)
rplA	S*	S*	S*	S*	S*

Table 3.4: Susceptibility tests of the selected mutants as compared to the wild-type according to CLSI standards. S: Sensitive, I: Intermediate, R: Resistant. *: the inhibition zone diameter of the mutant is significantly different than the WT at p.value ≤ 0.01. p.values > 0.01 are included between parentheses. Some mutants didn't shift their AMR class even when their inhibition zone diameter was significantly different than the wild-type.

As expected, not all the knockout mutants showed a different antimicrobial susceptibility phenotype than the wildtype, suggesting that they are not involved in antimicrobial resistance, or their involvement might be synergistic with other genes to be explored. Following antimicrobial susceptibility tests, *uhpB* and *mdaB* Keio mutants were resistant to ampicillin, whereas the wild type strain was susceptible (Table 3.4)

3.0.4 Conclusion

Identifying new genes associated to AMR or better characterizing the mechanisms associated to this phenomenon are relevant research topics, due to the increasing risk associated to AMR. For this purpose, we performed a systems analysis of *E. coli* interactome, through a network diffusion algorithm that, starting from known AMR-related genes reported in the Comprehensive Antibiotic Resistance Database (CARD) and in the PointFinder databases, identified novel putative genes associated to AMR. The network induced by the seed and the identified genes showed a community structure partly overlapping with known biological pathways (as annotated in KEGG [149] thus it was possible to associate part of the identified genes to known biological mechanisms, some of them known to be involved in AMR mechanisms and other not. We extracted a list of genes, prioritizing them by their relevance within the PPI network, and tested their corresponding knockout *E. coli* mutants with standard EUCAST/CLSI procedures for susceptibility to several antibiotics, and obtained an experimental validation for *uhpB*, *mdaB*, *rpmG* and *rplA*.

3.0.5 Discussion

Several studies highlighted the relevance of gene interaction study and its importance in AMR in different microbial pathogens [265]. Hence, we used network diffusion approach to examine several AMR mechanisms in *E. coli* to reveal new potential drug targets. Additionally relevant genes were validated though knockout experiments. In the present study, 127 genes were identified belonging to the following pathways: Ribosome (e.g. *rpmG* and *rplA*); Cationic antimicrobial peptide (CAMP) resistance (e.g. *acrAB*, *tolC*, *phoP* and *basR*); beta-Lactam resistance (e.g, *acrAB-tolC* and *ompC*); Fatty acid biosynthesis (*fab* genes); Peptidoglycan biosynthesis (*mur* genes); Fatty acid metabolism; Biotin metabolism.

Ribosome is a pathway already observed as enriched in other microorganisms namely *S. aureus*, *C. difficile*, *H. pylori* and *C. jejuni* [266, 267, 268, 269]. CAMP and beta-lactam resistance and peptidoglican biosynthesis were previously reported as relevant pathways in AMR mechanisms in *S. aureus*, *S. Typhi* and *E. coli O157:H7* [270]. Fatty acid biosynthesis was reported as significantly enriched in *E. coli O157:H7* [271]. Not surprisingly, genes related to efflux pumps (i.e. acrAB, tolC), were identified in this study as in previous ones [272]. By actively extruding antibiotics from the bacteria, multi-drug efflux pumps are under the lens of researchers since two decades as potential targets for novel drugs able to revert resistant phenotypes to several antibiotics [273]. New genes not previously identified by system biology approaches are *uhpB*, *uhpC* and *mdaB*. *uhpB* is a sensor histidine kinase of a two component system (TCS). TCSs have

been previously highlighted as primary pathways by which bacteria adapt to environmental stresses such as antibiotics. Knock out mutants of TCSs genes prioritized in the present study, namely, *phoP*, *cpxR*, and *basR* previously showed significant shift in their antimicrobial susceptibility reinforcing their role in AMR mechanisms [272].

Among the 127 genes identified, 13 were retained for validation experiments. Among these, *E. coli* knockout mutants of *uhpB*, *mdaB*, *rpmG*, and *rplA* showed a significant variation of their antimicrobial susceptibility (AST) in comparison to the wild-type, suggesting their functional involvement in related AMR mechanisms. In particular *uhpB* and *mdaB* knockout mutants showed a shifted AST against ampicillin, *rpmG* and *rplA* against ciprofloxacin, and *rplA* against streptomycin, suggesting their functional involvement in related resistant mechanisms.

Differently from previous findings, *uhpB* mutants were not fosfomycin resistant. The gene *uhpB* is an activator of the expression of *uhpT*, which is a phosphate inducible transporter responsible for the uptake of small molecules [229]. In the present study, results suggested the potential involvement of *uhpT* in the additional uptake of ampicillin. Regarding **mdaB**, further studies are needed to elucidate if this enzyme is able to inactivate ampicillin similarly to its detoxification role against quinones. As far as ciprofloxacin is concerned, knockout mutants *rpmG* and *rplA* were susceptible to this antimicrobial whereas the wildtype expressed an intermediate phenotype. *rpmG* is already known to be involved in DNA repair, its role in ciprofloxacin susceptible phenotype might be associated to the repair of DNA damages due to the inhibition of DNA synthesis by this antimicrobial agent. *rplA* is a ribosomal protein with no apparent connections with ciprofloxacin mode of action. *rplA* mutants were also susceptible to streptomycin whereas the wildtype was intermediate. Streptomycin mode of action relies on the inhibition of the protein synthesis. Although without a biological shift of the antimicrobial susceptibility, *rplA* mutants showed a significantly higher zone diameters than the wild type (p ¡ 0.01) for all 5 antibiotics suggesting *rplA* is involved in the mechanism of intrinsic resistance. This result is in line with Keio mutants MICs data reported by Liu and colleagues [274]. However, WT strain was found intermediate for streptomycin and ciprofloxacin in the present study whereas it was detected as sensitive for those antibiotics by the previous study [274]. Additional studies are needed to confirm and further investigate the potential role of *rplA* in intrinsic multidrug resistance or reduced susceptibility.

The four identified genes *uhpB*, *mdaB*, *rpmg* and *rplA* can be targeted either by a direct inhibition of their products (proteins) or indirectly by disrupting the regulatory pathway in which they are involved. For instance, *uhpB* gene is part of the two-component signal transduction system in *E. coli* responsible for sensing and transport of glucose-6-phosphate [275]. In this system, the protein *uhpC* senses the presence of glucose-6-phosphate, then, it interacts with *uhpB*. *uhpC-uhpB* interaction induces the activation of *uhpA* which, in turn, induces the expression of *uhpT* which is the protein responsible for the transport of many phosphorylated sugars including glucose-6-phosphate [276]. The susceptibility tests we carried out showed that *E. coli* becomes less sensitive to Ampicilin when *uhpB* is absent as the corresponding mutant AMR-profile switches from Sensitive (WT) to Intermediate. Therefore, *E. coli* resistance to Ampicilin can be mitigated through a therapeutic strategy that enhances *uhpB* activity. *rplA* is a risbosomal sub-unit. Ribosomes are responsible for the translation of mRNA into proteins [277]. We have seen that *E. coli* mutants lacking *rplA* are less resistant to Streptomycin and Ciprofloxacin than the wild type, therefore, therapies based on the direct inhibition of *rplA* protein or the inhibition of its expression would be effective against *E. coli* resistance to Streptomycin and Ciprofloxacin. *dskA* is an endogenous *E. coli's* protein that decrease the expression of *rplA*, so, a possible way to indirectly repress *rplA* would be by the administration of drugs structurally similar to *dskA*. The identification of gene candidates is a crucial step in drug discovery, however, the design of effective drugs is not straightforward due to the complexity and interconnectedness of biological pathways, in addition to the capability of the microbes to be resistant to multiple drugs [278]. Therefore, the findings of our work are a first step towards the development of new therapies to mitigate AMR in *E. coli* which require further studies

to identify the most effective strategy.

The ND approach was able to uncover ribosomes pathway, a pathway of interest targeted by different antimicrobial strategies, even though the network was not seeded with any ribosomal gene. This suggests that the outcomes of the ND gene prioritization are not limited by the initial seed genes. Most of the top-ranking genes belonged to known biological pathways: literature-based investigation and *in vitro* testing revealed implication in AMR for most of them. We performed susceptibility testing for thirteen selected mutants under nine different antibiotics: four mutants showed significant AMR shift as compared to the wild-type, in relation to five over nine tested antibiotics. Possibly the rest of the identified genes could be the false-positives, but it could be worth further *in vitro* testing using other antimicrobials.

3.0.6 Supplementary table

Table 3.5: The list of the 93 AMR-related genes prioritized by the network diffusion approach, in addition to the 34 seed genes.

Symbol	**b number**	**KEGG annotation**	**is_seed**	**diffusion_score_Sx**
uhpT	b3666	hexose-6-phosphate:phosphate antiporter	YES	0.801231429974594
uhpA	b3669	DNA-binding transcriptional activator UhpA	YES	0.794068832752411
glpT	b2240	sn-glycerol 3-phosphate:phosphate antiporter	YES	0.362348566172805
ptsI	b2416	PTS enzyme I	YES	0.308972861751571
acrA	b0463	multidrug efflux pump membrane fusion lipoprotein AcrA	YES	0.298002659860054
fabI	b1288	enoyl-[acyl-carrier-protein] reductase	YES	0.294422847975814
uhpB	b3668	sensory histidine kinase UhpB	NO	0.256489441518528
uhpC	b3667	inner membrane protein sensing glucose-6-phosphate	NO	0.251650558662633
basR	b4113	DNA-binding transcriptional dual regulator BasR	YES	0.233545958190829
folP	b3177	dihydropteroate synthase	YES	0.210048775401827
cyaA	b3806	adenylate cyclase	YES	0.18230290314585
pitB	b2987	metal phosphate:H(+) symporter PitB	NO	0.137596607913984
yhjB	b3520	putative DNA-binding transcriptional regulator YhjB	NO	0.130045062575375
ycaK	b0901	putative NAD(P)H-dependent oxidoreductase YcaK	NO	0.127905983952938

mdaB	b3028	NADPH:quinone oxidoreductase MdaB	NO	0.121634905014061
tufB	b3980	translation elongation factor Tu 2	YES	0.103510686077722
tufA	b3339	translation elongation factor Tu 1	YES	0.102895704838596
yeaE	b1781	methylglyoxal reductase YeaE	NO	0.06849965078668
mdtK	b1663	multidrug efflux pump MdtK	NO	0.066680419268121
emrD	b3673	multidrug efflux pump EmrD	NO	0.06280046585373
mprA	b2684	DNA-binding transcriptional repressor MprA	NO	0.059638151998981
pmrR	b4703	putative bitopic inner membrane protein	NO	0.059283295703077
plsX	b1090	putative phosphate acyltransferase	NO	0.045645668710855
marB	b1532	multiple antibiotic resistance protein MarB	NO	0.042808523055997
yeaD	b1780	putative aldose 1-epimerase YeaD	NO	0.042805044003558
macB	b0879	ABC-type tripartite efflux pump ATP binding/membrane subunit	NO	0.042568222182127
pagP	b0622	Lipid A palmitoyltransferase	NO	0.042435308529878
fabZ	b0180	3-hydroxy-acyl-[acyl-carrier-protein] dehydratase	NO	0.040334558855965
lpxC	b0096	UDP-3-O-acyl-N-acetylglucosamine deacetylase	NO	0.038514304261013
phoP	b1130	DNA-binding transcriptional dual regulator PhoP	NO	0.036996396848245
murC	b0091	UDP-N-acetylmuramate–L-alanine ligase	NO	0.036996359525686
murD	b0088	UDP-N-acetylmuramoyl-L-alanine–D-glutamate ligase	NO	0.036076039129444
plsY	b3059	putative glycerol-3-phosphate acyltransferase	NO	0.035381713892495
sodA	b3908	superoxide dismutase (Mn)	NO	0.035229027003966

murG	b0090	N-acetylglucosaminyl transferase	NO	0.034981285821487
lpxA	b0181	acyl-[acyl-carrier-protein]–UDP-N-acetylglucosamine O-acyltransferase	NO	0.034868411972527
mrdB	b0634	SEDS family protein MrdB	NO	0.034829015158298
glpK	b3926	glycerol kinase	NO	0.034114114611957
skp	b0178	periplasmic chaperone Skp	NO	0.033238662065734
glpF	b3927	glycerol facilitator	NO	0.03310588034669
murF	b0086	D-alanyl-D-alanine-adding enzyme	NO	0.032869303546023
topA	b1274	DNA topoisomerase 1	NO	0.032665842180443
glmU	b3730	fused N-acetylglucosamine-1-phosphate uridyltransferase and glucosamine-1-phosphate acetyltransferase	NO	0.032385254520689
folB	b3058	dihydroneopterin aldolase	NO	0.032346391278466
mreB	b3251	dynamic cytoskeletal protein MreB	NO	0.032166095956974
bamB	b2512	outer membrane protein assembly factor BamB	NO	0.032023279208861
mraZ	b0081	DNA-binding transcriptional repressor MraZ	NO	0.031774983867738
accD	b2316	acetyl-CoA carboxyltransferase subunit beta	NO	0.031542084511844
yceD	b1088	DUF177 domain-containing protein YceD	NO	0.030150344623432
ftsA	b0094	cell division protein FtsA	NO	0.02925681968476
ftsQ	b0093	cell division protein FtsQ	NO	0.028884357036002
dnaG	b3066	DNA primase	NO	0.027961243041426
rpoH	b3461	RNA polymerase, sigma 32 (sigma H) factor	NO	0.027836259256324
rpmB	b3637	50S ribosomal subunit protein L28	NO	0.026811827074246
rpoC	b3988	RNA polymerase subunit beta'	NO	0.026676178298001
rpoD	b3067	RNA polymerase, sigma 70 (sigma D) factor	NO	0.026580497011964

rpoA	b3295	RNA polymerase subunit alpha	NO	0.026575937057303
rpsJ	b3321	30S ribosomal subunit protein S10	NO	0.02624190029595
rplQ	b3294	50S ribosomal subunit protein L17	NO	0.026170361371692
rpsA	b0911	30S ribosomal subunit protein S1	NO	0.02607704673147
rpmI	b1717	50S ribosomal subunit protein L35	NO	0.026002415434723
rpsR	b4202	30S ribosomal subunit protein S18	NO	0.025985678836753
rplA	b3984	50S ribosomal subunit protein L1	NO	0.0259136634777
rpsM	b3298	30S ribosomal subunit protein S13	NO	0.025779235397691
rpsQ	b3311	30S ribosomal subunit protein S17	NO	0.025716648865321
rplJ	b3985	50S ribosomal subunit protein L10	NO	0.025674675532255
rpsB	b0169	30S ribosomal subunit protein S2	NO	0.025624614185735
rpsF	b4200	30S ribosomal subunit protein S6	NO	0.025615165161537
priB	b4201	primosomal replication protein N	NO	0.025534441866644
rpsC	b3314	30S ribosomal subunit protein S3	NO	0.025494340114023
metG	b2114	methionine–tRNA ligase	NO	0.025396077083455
rplW	b3318	50S ribosomal subunit protein L23	NO	0.025352534812794
rplV	b3315	50S ribosomal subunit protein L22	NO	0.025326888492826
rplM	b3231	50S ribosomal subunit protein L13	NO	0.02512905531526
arfA	b4550	alternative ribosome-rescue factor A	NO	0.02494438702434
infA	b0884	translation initiation factor IF-1	NO	0.024396031467337
rmf	b0953	ribosome modulation factor	NO	0.024340506076583
nusG	b3982	transcription termination factor NusG	NO	0.024002680550944
hpf	b3203	ribosome hibernation-promoting factor	NO	0.023081728448611
mipA	b1782	MltA-interacting protein	YES	0.645942626575027
fabG	b1093	3-oxoacyl-[acyl-carrier-protein] reductase FabG	YES	0.34234427867348
basS	b4112	sensor histidine kinase BasS	YES	0.235592389895031
nemA	b1650	N-ethylmaleimide reductase	NO	0.14475594834719

ftsI	b0084	peptidoglycan DD-transpeptidase FtsI	YES	0.142851128013778
yieF	b3713	chromate reductase	NO	0.108647681207274
acrD	b2470	multidrug efflux pump RND permease AcrD	NO	0.069119537541059
fabD	b1092	[acyl-carrier-protein] S-malonyltransferase	NO	0.046859905444703
fabB	b2323	3-oxoacyl-[acyl carrier protein] synthase 1	NO	0.042255646147304
mrdA	b0635	peptidoglycan DD-transpeptidase MrdA	NO	0.036960250745843
katG	b3942	catalase/hydroperoxidase HPI	NO	0.036943599693355
ftsK	b0890	cell division DNA translocase FtsK	NO	0.032130874794963
uvrB	b0779	excision nuclease subunit B	NO	0.030028510254639
rpmG	b3636	50S ribosomal subunit protein L33	NO	0.026232132734047
secY	b3300	Sec translocon subunit SecY	NO	0.026123608299064
lepA	b2569	30S ribosomal subunit biogenesis factor LepA	NO	0.025071492393252
nfsB	b0578	NAD(P)H nitroreductase NfsB	YES	0.530049625058895
soxR	b4063	DNA-binding transcriptional dual regulator SoxR	YES	0.367411944018482
cpxR	b3912	DNA-binding transcriptional dual regulator CpxR	YES	0.282458400582228
lamB	b4036	maltose outer membrane channel/phage lambda receptor protein	YES	0.147295305785713
pmrD	b2259	signal transduction protein PmrD	NO	0.048886958742469
ompC	b2215	outer membrane porin C	NO	0.038869264970844
ompT	b0565	DLP12 prophage; protease 7	NO	0.033359062142597
ftsZ	b0095	cell division protein FtsZ	NO	0.02930555460478
secA	b0098	protein translocation ATPase	NO	0.028775823923588
ompA	b0957	outer membrane protein A	NO	0.026970200813083
nfsA	b0851	NADPH-dependent nitroreductase NfsA	YES	0.545080241576949
murA	b3189	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	YES	0.250962834085843

oxyR	b3961	DNA-binding transcriptional dual regulator OxyR	NO	0.037278135411326
crp	b3357	DNA-binding transcriptional dual regulator CRP	NO	0.03542244755828
lacZ	b0344	beta-galactosidase	NO	0.034991121532641
acrR	b0464	DNA-binding transcriptional repressor AcrR	YES	0.473670373698583
parE	b3030	DNA topoisomerase IV subunit B	YES	0.23843747364297
rpoB	b3987	RNA polymerase subunit beta	YES	0.079002644545744
emrE	b0543	DLP12 prophage; multidrug/betaine/choline efflux transporter EmrE	YES	1.70944372066826
gyrB	b3699	DNA gyrase subunit B	YES	0.134912375109366
mdfA	b0842	multidrug efflux pump MdfA/Na(+):H(+) antiporter/K(+):H(+) antiporter	YES	0.300743537765386
rob	b4396	DNA-binding transcriptional dual regulator Rob	YES	0.548635237479999
araC	b0064	DNA-binding transcriptional dual regulator AraC	NO	0.055465134711617
soxS	b4062	DNA-binding transcriptional dual regulator SoxS	YES	0.355662537069716
ompF	b0929	outer membrane porin F	YES	0.199356633775028
marR	b1530	DNA-binding transcriptional repressor MarR	YES	0.395965728916185
parC	b3019	DNA topoisomerase IV subunit A	YES	0.335664392645696
acrB	b0462	multidrug efflux pump RND permease AcrB	NO	0.065828860622623
tolC	b3035	outer membrane channel TolC	NO	0.086053853250199
marA	b1531	DNA-binding transcriptional dual regulator MarA	NO	0.105650761716941
gyrA	b2231	DNA gyrase subunit A	YES	0.154388786749611
ampC	b4150	beta-lactamase	YES	0.291891226132292

Chapter 4

Context-Specific Genome-Scale Constrained Models Using Transcriptomics, Flux Variability, and Network Topology

4.1 Introduction

Living cells are equipped with the necessary machinery to accomplish their biological functions. This machinery can be modeled as networks of different molecules interacting together to generate a given phenotype. There are three major biological networks: 1) Regulatory networks, 2) Transduction networks and 3) Metabolic networks. Metabolic networks are of great interest because they provide all biological processes with energy and elemental building molecules such as amino acids, fatty acids and nucleic acids. Due to their important role, studying metabolic networks can shed light onto many regulatory processes that control the cell's functions.

Metabolic networks are large and complex as they involve many interconnected reactions which convert many shared metabolites, therefore, comprehensive representation of such networks is challenging. Genome Scale Metabolic Modeling (GSMM) is an effective approach used to model the metabolic potential of cells by combining high-throughput genomics data and prior biochemical knowledge [279]. A multitude of curated GSMMs of human cells, bacteria and yeasts are available in different databases such as BIGG [107], KEGG [108] and ModelSEED [111]. To make use of these GSMMs, one can use constraint-based modeling (CBM), i.e, the combination of constraint-based methods such as Flux Balance Analysis (FBA) and GSMM to model the metabolic properties of the cells and predict their outcomes [112]. FBA is a constraint-based methods as it relies on: 1) The mass balance constraint arising from the assumption of a steady-state, i.e, the equality of biomolecules' production and consumption rates; 2) The capacity constraints that define the lower and upper bounds imposed to the fluxes of the reactions.

As the number of published GSMMs is growing over the past years, CBM could be used in different context to study animal and microbial metabolism [280]. In the field of food biotechnology, CBM has been used to improve microbial yields of metabolites of interest such as low-calorie sweeteners, vitamins and bioactive peptides [281, 282], as well as the design of microbial strains able to improve certain food nutritional and safety aspects such as enhancing shelf-lives and the biotransformation of indigestible compounds into edible food products [283].

Given the adaptive nature of the living beings, the cells are able to modulate their metabolism according to the environment in which they live and grow. For instance, in the presence of glucose, *Saccharomyces cerevisiae* produces the energy required for its growth predominantly by glycolysis. When glucose becomes limiting, the yeast undergoes a diauxic shift allowing it to produce energy from ethanol. To do so, the yeast rewires its regulatory networks, mainly, the central energy signaling pathways [284]. This implies that, in each specific context, not all the metabolic network is operating

but only a subset of specific reactions that are relevant to achieve the objective biological function. Therefore, to accurately model the cells' metabolism, there is a need to focus on the active reactions within the GSMM and discard the non-relevant ones. Different methods have been developed to create these context-specific GSMMs. Some of them use experimental data such as transcriptomics to define the active/inactive genes/reactions, such as the Gene Inactivity Moderated by Metabolism and Expression (GIMME) [285], and the FASTCORE family algorithms which determines the core (active) reactions in a given condition basing on bibliography or proteomics data [286]. Gene expression is strongly associated and influences metabolic networks since most of the biochemical reactions occurring inside cells are gene-regulated. Genes encode proteins, and proteins can have enzymatic activity involved in the cell metabolism. Thus mRNA expression is a proxy to measure enzyme (protein) levels inside the cells to trigger metabolic reactions. Other GSMM reduction methods are purely computational, i.e. they perform the reduction of the GSMM based only on its structure. Ataman et al. [287] developed redGEM which is a bottom-up reduction approach, contrary to top-down approaches such as GIMME, redGEM doesn't start from the whole GSMM, but starts by the construction of a small and basic core network, which is then extended to obtain a reduced GSMM representing the active reactions in the context that is defined by the user. Jonnalagadda et al. [288] integrated graph-theory into their GSMM reduction algorithm. To define the essential reaction to be included in the reduced GSMM, they depict the network structure of the GSMM as a bipartite graph connecting reactions and their metabolites. They analyze the network to find the reactional routes that lead to the production of the same metabolites, then, out of these redundant reactions, they keep the ones that optimize better the objective function.

Another consequence of reactions' redundancy is that objective functions, e.g. the growth, can be attained through different routes. To explore the possible (sub)optimal reaction fluxes, Mahadevan et al. [124] have proposed a variant of FBA so called Flux Variability Analysis (FVA). It's an effective tool to investigate the distribution of the reactions fluxes which can shed light on alternative and interesting regulatory pathways.

In addition of being context-specific and redundant, metabolism is also dynamic. Therefore, classic FBA which assumes a static state of the network may not be the most suitable to analyze GSMMs. Mahadevan et al. [126] proposed Dynamic FBA (dFBA) to account for the evolving nature of metabolism and define phase or time-point-specific fluxes.

Objective

In this work, we aimed at improving GSMM reduction by developing a new method taking into account both experimental and network-topology information. Conventional reduction methods require experimental data with several time-points - which is often hard to acquire - to build reduced GSMMs able to model the systems' dynamics. To overcome this limitation, we added another layer of information to our method, i.e. the network-topology-based information which contribute to obtain reduced GSMMs with more relevant context-specific reactions.

Summary

Starting from time-series transcriptomics and metabolomics data, as well as the GSMM of the yeast strain *Saccharomyces cerevisiae T73*, we developed and calibrated a new GSMM-reduction algorithm. The method relies on the combination of constraint-based modeling (dFVA), experimental data (transcriptomics and metabolomics) as well as network-based approaches (centrality measures) to reduce the GSMM by discarding non-relevant reactions and keep only the reactions that are more likely to be active in the specific experimental context.

We found that the synergetic combination of different centrality measures could effectively define the relevant reactions to keep in the reduced GSMM.

The method was compared to the conventional GSMM-reduction methods GIMME [285] and FASTCORE [286] in terms of the generated reduced GSMMs and prediction performances. The corresponding results are not reported in this chapter[1].

4.2 Materials and Methods

Experimental data

In this work, we used data related to the yeast strains *Saccharomyces cerevisiae T73* comprising its Genome Scale Metabolic Model, longitudinal transcriptomics data and extracellular metabolites.

We measured in ten sampling times extracellular metabolites, including sugars, organic acids, main fermentative by-products and yeast assimilable nitrogen (YAN) in the form of amino acids and ammonia, following the experimental protocol defined in [289]. We also determined the concentrations of higher alcohols and esters for each sampling time. Volatile compounds extraction and gas chromatography were performed following the protocol by Rojas et al. [290]. Physiological and biomass parameters, including OD600, dry weight (DW), colonies-forming unit (CFUs), and average cell diameter (ACD), were determined at each sample time, provided that the cell sample was sufficient to perform the corresponding measure.

Transcriptomics analysis was conducted on cells harvested from the fermentation broth of each biological triplicate at three different time points. The time points were during the growth phase (T1: 20 hours), at the end of the growth phase (T2: 26.5 hours), and early stationary phase (T3: 43.25 hours. To obtain the cells, the broth volume was collected from the reactor and transferred to a polypropylene tube, which was then centrifuged (4.000 rpm, 5 minutes, 4ºC) to pellet the cells. The supernatant was discarded, and the tube was flash-frozen in liquid nitrogen and stored at -80ºC until total RNA extraction. Following the manufacturer's protocol, total RNA was extracted using the High Pure RNA Isolation Kit (Roche, Mannheim, Germany). The samples were sequenced using the Illumina Hiseq 2000, paired-end reads 75 bases long and were deposited under the BioProject ID PRJNA473087. The sequence reads were trimmed and quality filtered using Sickle (minimum read length of 50, minimum quality per base of 23) and then aligned to the strain genomes using bowtie2 [291]. Gene counts were obtained using HTSeq-count version 0.9.0.[292].

The metabolic data (to be provided when the work is published) were used to define the flux's bounds, whereas transcriptomics data were combined to the network-topology measures to distinguish between relevant and non-relevant reactions (next sections).

Dynamic Flux Variability Analysis dFVA to define the time-dependent relevant reaction sets

Through their different growth phases, yeasts modulate their metabolism to attain the objective functions at each phase. These objective functions can be achieved through different and redundant reaction routes. This results in different patterns of the reactions' flux distribution [124].

To measure the possible reactions' fluxes at each transient state, we carried out a dynamic Flux Variability Analysis, i.e, at each time point t, the maximums and minimums of the reactions fluxes v maximizing the growth were computed as follows:

$$\max v_i \quad \text{subject to} \quad C,\ v_{\min} \leq v \leq v_{\max},\ Z = c^T \cdot v$$

$$\min v_i \quad \text{subject to} \quad S \cdot v = 0,\ v_{\min} \leq v \leq v_{\max},\ Z = c^T \cdot v$$

[1]This work is an integral part of PhD. Diego Troitiño's thesis, from the Bio2Eng group headed by Prof. Eva Balsa-Canto, therefore, some parts can't be described in detail, mainly, the newly developed scoring functions, but we do our best to clearly explain the methods, mainly, our part of collaboration to this project. The work is being finalized and some results are still to be generated. The final paper will be submitted for publication soon.

where **S** represent the stoichiometric matrix of the metabolic network, and **v** is a column vector of the fluxes of each reaction. v_{min} and v_{max} are the lower and upper bounds of the flux v, respectively. We remark that fluxes v are a combination of the original metabolic compounds together with the kinetic constants of the considered reactions: Flux Balance Analysis is a simplification with respect to the full set of differential equations describing the system of metabolic reactions in a cell, but it allows to find a solution as a linear system of equations under the stationarity assumption $S \cdot v = 0$. Z is the maximum growth value, a scalar objective function obtained as a weighted sum of fluxes, with c the objective function weight vector for each flux.

Networks construction

In this work we tried to estimate the relevance of each metabolic reaction by using information contained in two related networks: 1) a gene co-expression network derived from experimental observations (gene expression profiles at the different time points of the experiment); 2) the biochemical reaction network as extracted from the publicly available curated yeast GSMM.

Gene co-expression network

The data-driven gene co-expression network was built using *WGCNA* R package [171] from the Next-Gen sequencing transcriptomics data, consisting of 6169 genes and nine samples. First, the data were filtered out of lowly expressed genes (< 5 reads in 3 samples) then normalized using TMM method [146]. The processed and normalized data contained 5480 genes. Secondly, we computed the the Spearman correlation [170] between gene pairs and retained only the significant correlations at p.value ≤ 0.05. As a result, we obtained an edge list of 2446270 rows which represent a weighted gene-gene network of 5476 nodes and 2446270 weighted edges. Four genes (5480-5476) were not significantly correlated to any other gene, so they were not included in the network. The correlation values were converted into network weights by considering their absolute values (only positive weights) thus considering correlation and anti-correlation as an indicator of strong relationship.

This pure data-driven approach can result in many spurious interactions that may not actually exist between certain genes. To filter out such interactions, we matched the data-driven network with the yeast's protein network available on STRING database [189] that thus provided the structural backbone of the network, using the correlation values as weights. Every weight referring to an interaction that was not reported in STRING was omitted. As a result, we got a filtered and weighted gene-gene (or protein-protein) network of 5091 nodes and 169842 edges.

Reaction networks

Genes expressed by the yeast are not all involved in metabolic pathways. In this second procedure, we focused on the yeast's metabolic pathways reported in the Genome Scale Metabolic Models (GSMM) model. GSMMs are encoded according to the Systems Biology Markup Language (SBML) standards developed for the communication and storage of computational biological models [293]. GSMM can be saved in human-readable files such as .xml and .json. In this work, the yeast's GSMM was saved in .xml file which mainly contains the stoichiometric matrix of the metabolite-reaction network in addition to the mass equation.

The metabolic network is thus represented by the stoichiometric matrix. It can be interpreted as a bipartite network where the nodes are of two types: Reactions and Metabolites, and the links exist only between the nodes of different types, i.e. between the reactions and their associated metabolites. As we were interested in identifying (non)relevant reactions, we converted the bipartite network into a unipartite or one-mode reaction network. To do so, we first removed the following pool metabolites: ATP, ADP, NAD, NADH, NADP, NADPH, O_2, H_2O, P_i, H^+, Glu_L and CO_2. These metabolites are consumed and produced by a large number of reactions, therefore, suppressing them is crucial to avoid distorting the topological proprieties and over-connecting the network due to the linkage of functionally unrelated reactions. Then,

we created edges (links) between pairs of reactions that share at least one metabolite. The conversion of the reaction-metabolite network into a reaction-reaction network is done by multiplying the stoichiometric matrix by its transpose:

$$\text{Reaction-Metabolite network} = S$$

$$\text{Reaction-Reaction network} = S.S^T$$

where S is the stoichiometric matrix and S^T its transpose. The stoichiometric matrix can be easily extracted from the .xml files, otherwise, function such as *makeRactionNetwork()* and *bipartite_projection()* from the *NetPathMiner* [294] and *igraph* [295] R packages can be used for automatic parsing and conversion of GSMMs into reaction networks starting from .xml files.

Following the steps described above, we built a reaction network with 3433 nodes and 5978 edges.

Centrality measures of the reactions

Centrality measures (CMs) are scores used to define the importance and the influence of a node within a network (see Introduction and Chapter 3). In this work, we tested several CMs in order to select the most relevant reactions. We considered local, semi-local and global CMs:

- **Local CMs:** reflect the node's importance within its first neighborhood such as Degree, Cluster_rank and Clustering coefficient CMs.
- **Semi-Local CMs:** take into account an environment beyond the node's first neighborhood but not the global network, such as H-index, Local H-index and Neighborhood connectivity.
- **Global CMs:** measures the node's importance by taking into account the entire network, such as Collective_influence, Closenesses, Eccentricity and Betweenness CMs.

Most of the mentioned CMs are already described in Chapter 1. Here we describe the newly introduced CMs: Closeness and Eccentricity.

- **Closeness CM**: measures the distance (shortest path) of a node to the other nodes in the network [45]. It represents the geographic position of the node in the network. Nodes having high closeness spreads information fast as they are close to many nodes:

  $$\text{Closeness(v)} = \frac{1}{\sum_{t \in V \setminus \{v\}} \text{dist}(v,t)}$$

  where $\text{dist}(v,t)$ is the distance between the nodes v and t.

- **Eccentricity CM:** also called "Harary Graph Centrality" or "Jordan Centrality" [296]. It is similar to Closeness CM, but it takes into account the longest of the shortest paths between a given node and the other nodes in the network:

  $$\text{Eccentricity(v)} = \frac{1}{\max\{\text{dist}(t,v) : t \in V\}}$$

As each CM captures a specific property of the node, taking into account the CMs individually may under/overestimate the node's importance [30]. Therefore, we also considered combining several CMs to create a comprehensive score that can reflect better the node's influence. We computed the Integrated Value of Influence score (IVI) available on the *influential* R package [297]. IVI score combines the most important centrality measures, namely, Degree Centrality (DC), Betweenness (BC), ClusterRank

(CC), Neighborhood connectivity (NC), Local H index (LHC) and Collective influence (CIC) in a way that synergizes their effect to identify influential nodes in the network in an unbiased way. The IVI score of each node i is formulated by the authors as follows:

$$\mathrm{IVI}_i = (\mathrm{DC}_i + \mathrm{LHC}_i)\left((\mathrm{NC}_i + \mathrm{CC}_i)\left(\mathrm{BC}_i + \mathrm{CIC}_i\right)\right)$$

with:

$$DC(v_i) = \mathrm{N}(v_i)$$

where $\mathrm{N}(v_i)$ is the set of direct neighbors of node v_i.

$$BC(v_i) = \sum_{a \neq v_i \neq b} \frac{\sigma(a, b|v_i)}{\sigma(a, b)}$$

where $\sigma(a, b)$ is the number of shortest paths from a to b, and $\sigma(a, b|v_i)$ is the number of those paths passing through v_i

$$NC(v_i) = \frac{1}{\mathrm{N}(v_i)} \sum_{v_j \in \mathrm{N}(v_i)} \mathrm{DC}(v_j)$$

where v_j is a node from the direct neighborhood of v_i

$$CIC(v_i) = [DC(v_i) - 1] \sum_{v_j \in \partial Ball(v_i, d)} [DC(v_j) - 1]$$

$Ball(v_i, d)$ is the set of nodes inside a ball of radius d (shortest path) around v_i, and $\partial Ball(v_i, d)$ is the frontier of the ball

$$CC(v_i) = \frac{DC_i}{\sum_{j \in N(i)} DC_j + DC_i} C(v_i)$$

where $C(v_i)$ is the clustering coefficient of node v_i:

$$C(v_i) = \frac{2 \times e_i}{DC_i(DC_i - 1)}$$

where e_i is the number of edges between the neighbors of node v_i.

$$H_{index}(v_i) = \max\{h : N_h(v_i) \geq h\}$$

with $N_h(v_i)$ is the number of neighbors of v_i with a degree $\geq h$.

$$LHC(v_i) = H_{index}(v_i) + \sum_{v_j \in N(v_i)} H_{index}(v_j)$$

The effectiveness of IVI score relies on the fact that it accounts for several node's centralities in different environments: 1) the local influence of the node, i.e. within its direct neighborhood (DC and CC); the semi-local node's influence, i.e. in the environment beyond its direct neighborhood (LHC and NC); 3) the node's influence on the entire network (BC and CIC).

For each node in the data-driven gene co-expression network and in the reaction network extracted from the GSMM, we computed the aforementioned CMs singularly in addition to IVI scores.

Susceptible-Infected-Recovered model: another way to measure nodes' importance

The Susceptible-Infected-Recovered (SIR) model is mathematical model involving ordinary differential equations used in epidemics to measure the spread of a disease in a population [298, 299]. In our case, the population is the reactions network, and each reaction is an individual. SIR-based method assumes that at t_0 one random reaction is "Infected" (I) while all the other reactions are "Susceptible" (S), then, starting from the "Infected" node the disease is propagated through the network. The "Susceptible"

nodes become "Infected" with a rate which depends on their "Infected" neighbors. Finally, the algorithm stops when there are no more "Infected" individuals, i.e, all nodes are in state "Susceptible" or "Recovered" (R) [300]:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI,\\ \frac{dI}{dt} &= \beta SI - \gamma I,\\ \frac{dR}{dt} &= \gamma I.\end{aligned}$$

β is the disease transmission rate, it represents the probability for an infected node to transmit "disease" to its susceptible neighbors at time t and γ is the recovery rate, it represents the probability of an "Infected" node to become "Recovered".

The SIR-based method is used to rank reactions according to their importance, that is, their role in spreading the "disease" as follows:

1. The "disease" spread, measured by the number of "Recovered" nodes, is computed on the original reaction network using SIR model.
2. The "disease" spread is recomputed on a perturbed network after the removal of one reaction.
3. A SIR score is defined by computing the difference between the "disease" spread in the original and perturbed reaction networks.
4. Reactions resulting in a high SIR score are highly ranked and considered to be more important, as their removal induces a significant variation of the disease spread.

The steps above are repeated until all of the reactions have been removed from the network one time, and involved in the network $k-1$ times, where k is the number of nodes in the original reaction network. The approach described here is a combination of SIR model and Leave-One-Out method. It's implemented in the *sirir()* function from the *influential* R package [297].

Selection of relevant reactions

To reduce the GSMM and keep the most relevant reactions, we combined dFVA results, transcriptomics, CMs, IVI and SIR scores as follows:

- **Step 1:** the fluxes computed by dFVA were used to create a score allowing to classify the reactions into two sets: relevant and non-relevant reactions.
- **Step 2:** among the non-relevant reaction set, we wanted to recover other reactions basing on their transcriptomics expression level. However, the expression level of a gene isn't always correlated with its actual function, i.e. some genes can be lowly expressed but still play an important role in the metabolic network, and vice versa. To cope with this issue, we created a score that complements the transcriptomics information by taking into account also the network-topology-based information captured by the different CMs, the IVI score as well as the SIR score. Practically, for each non-relevant reaction defined by the dFVA-based approach, a new score is computed. This score takes into account the reaction's transcription level in addition to one of the following metrics:
  - individual CMs, e.g, degree, or betweenness...
  - The average value of all CMs.
  - SIR score.
  - IVI score.

The reactions to be recovered from the Non-relevant set and moved to the Relevant set must obtain a score above a certain threshold. This score is the combination of transcriptomics-based information (number of reads) and the network-topology-based information (CMs, IVI and SIR scores). The resulting context-specific GSMM will then be composed of relevant reactions selected by the dFVA-based approach as well as reactions recovered by the approach combining transcriptomics and network topology. The reactions that couldn't be recovered from the non-relevant reaction set were those which had both a low expression level and a low network-topology-based score.

Note that the data-driven gene co-expression network can't be directly used to measure the reactions importance as each metabolic reaction can be either regulated by one or more genes, or not regulated such as spontaneous and translocation reactions. The network-topology-based scores (i.e. CMs, IVI, SIR) assigned to each reaction represent the average scores of the gene(s) regulating that reaction. Non gene-regulated reactions were discarded.

Consistency of the reduced GSMMs

The relevant reactions selected based on the gene co-expression network as well as those selected based on the reaction network directly extracted from the yeast GSMM, were used to construct the reduced and context-specific yeast GSMMs. The reduced GSMMs were tested for their biological consistency, i.e, their ability to grow, to produce CO_2 and sufficiently produce ATP to maintain cell survival.

Algorithm implementation

The steps previously described were implemented in our algorithm called Genome-scale model assembly by Network flux variability, Expression and network Topology (GeNETop). The different constraint-based modeling operations were carried out using the *AMIGO2* [301] and *COBRA tool boxes* [302, 303], whereas the different network approaches were carried out using *WGCNA*, *igraph* and *influential* R packages [171, 295, 297].

4.3 Results

At the moment of writing this section, this work is still being finalized, so, we'll describe the primary results obtained so far.

The scores used to select the relevant reactions were derived from and calculated on two available networks: 1) The data-driven gene co-expression network; 2) The curated reaction network extracted from the original GSMM. These scores were combined with the reactions' expression levels to recover reactions from a previously non-relevant reaction set defined by dFVA, in order to obtain a consistent reduced yeast's GSMM.

Table 4.1 reports the biological consistency of the reduced GSMM obtained by the two approaches. We can see that the scores derived from the data-driven correlation network failed to obtain a reduced and biologically consistent GSMM, regardless the network-topology-based metric that was used. The obtained reduced models were not able to grow and produce CO_2 and ATP. In addition, the primary metabolic pathways were incoherent and disconnected. Similarly, individual CMs and SIR scores computed on the curated reaction network resulted also in biologically non consistent reduced GSMM. However, averaging several CMs partially improved the reduced GSMM, i.e, the primary metabolism was recovered even if GABA pathway and secondary metabolism were not consistent. Finally, despite some lacks in the secondary metabolism, the synergetic combination of CMs using IVI score got the best results in terms of consistency as it could recover even the low-level metabolic flux reactions.

	Individual CMs	Averaged CMs	SIR	IVI
Data-driven network	Non consistent	Non consistent	Non consistent	Non consistent
Curated reaction network	Non consistent	Partially consistent	Non consistent	Consistent with some lacks in secondary metabolism

Table 4.1: Biological consistency of the reduced GSMM based on the scores derived from the gene co-expression network (Data-driven network) and reaction network of the original GSMM (Curated reaction network). CM: Centrality Measure. SIR: the score computed by the Susceptible-Infected-Recovered model. IVI: the Integrated Value of Influence score.

We were interested in the different CMs calculated on the GSMM reaction network, mainly those used to compute the IVI scores. Figure 4.1 shows that the distribution of IVI scores and the CMs used to compute it correspond to a scale-free topology of the network, i.e, there are few nodes (reactions) with high centrality values and the majority of the nodes have low CMs (left-skewed distribution), and this is a characteristic of many biological networks such as gene regulatory networks, protein networks and metabolic networks,[304]. Secondly, except for Cluster Rank CM,the IVI score correlates well with the individual CMs suggesting that the IVI function is able to capture in a simultaneous way the information provided by each individual CMs. We also notice that some individual CMs are correlated. The highest correlation (0.93) is between Degree and Local H-Index CMs, implying that these CMs hold the same information, i.e. using the one or the other individually may lead to the same ranking of the reactions. Whereas, the lowest correlation (0.062) is obtained for Betweenness and Neighborhood connectivity CMs. Out of the 15 pairwise correlation measured between the individual CMs, 6 were strong (> 0.6): Degree-Local H-Index (0.919), Degree-Betweenness (0.748), Degree-Collective Influence (0.852), Betweenness-Local H-Index (0.638), Collective Influence-Local H-Index (0.928) and Collective Influence-Betweenness (0.626). Despite these pairwise correlations detected for some CM pairs, their combination through the IVI function led to the most biologically consistent reduced GSMM.

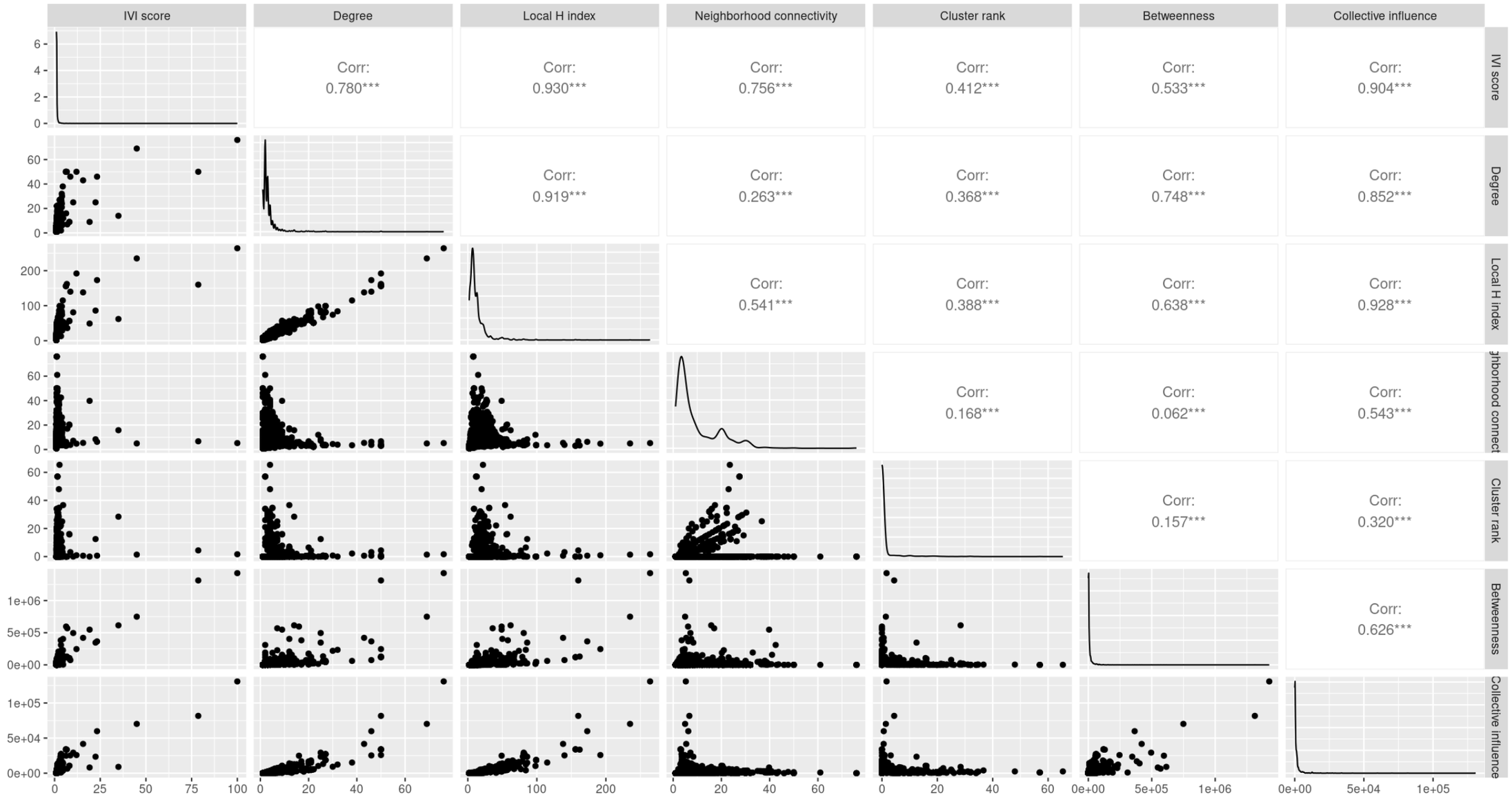


Figure 4.1: The distribution of the reaction's IVI scores and centrality measures and their correlation measured through Spearman metric. The centrality measures are those computed from the GSMM reaction network.

4.4 Conclusion

Through this analysis, aiming at identifying the core metabolic reactions in yeast cells undergoing an observational experiment, we could see that the type of network centrality metrics as well as the types of networks used to compute them have a significant impact on the biological relevance of the metabolic reactions selected to build a reduced and context-specific GSMM. Data-driven networks inferred from experimental transcriptomics data failed to achieve consistent results even though they were refined by matching them with curated gene-gene (protein-protein) network from the STRING database [189]. On the other hand, the reaction network from the curated yeast's GSMM allowed obtaining scores that were useful for defining the relevant reactions. However, not all scores were effective, but only the one that combines several CMs either by averaging them or combining them according to the IVI function.

The results of our work demonstrate the robustness and the consistency of manually curated networks as compared to inferred ones. Furthermore, they stress the need for choosing the adequate analysis methods and metrics to extract relevant information from these networks.

4.5 Discussion

It's important to stress that taking into account each CM separately can result in an underestimation or overestimation of the nodes' importance because some of them can have high global centrality but low local centrality or vice versa. As an example taken in a different context, on August 27, 2024, a landslide occurred in Maurienne train station in France [305], which caused the interruption of train traffic between Milan and Paris. Despite the low connectivity (degree) of Maurienne station in the transportation network, as compared to highly connected hubs like Paris and Milan, it has a high betweenness as its congestion led to significant disruption in the Franco-Italian train network. Thus, the importance of Maurienne station could be underestimated if only its degree CM was considered and not also its betweenness CM.

In metabolic networks, it has been shown that the correlations between reactions/metabolites essentiality and their degree centrality (connectivity) was weak, as reactions that are essential for growth have often a unique route (Elementary flux Modes) which involve poorly connected but essential enzymes and metabolites [306, 307]. On the other hand, PageRank CM [308] was combined with reactions' fluxes to create a score allowing the selection of metabolites biomarkers related to diabetes from human GSMM [308]. Furthermore, Beguerisse-Díaz et al. [309] have studied the metabolic changes between healthy and diseased hepatic cells affected by the Primary Hyperoxaluria Type 1. They found that reactions related to diseased state didn't exhibit a significant variation in their fluxes, yet they showed large changes in their PageRank centrality.

These findings demonstrate that: 1) flux-balance-based models to used describe cell metabolic activity can be complemented by network-based analyses, mainly, the use of CMs to identify key reactions and metabolites; 2) CMs capture different types of information so they have to be chosen adequately in order to extract relevant information from the network; 3) a comprehensive centrality score combining several CMs can provide a better estimation of nodes' importance.

Several works have addressed this problem to try to find the "best" combination of CMs that can hold as much information as possible about the nodes. Chua et al. [310] have created UniScore, a score that takes into consideration five different CMs to predict essential proteins in protein-protein interaction networks. They evaluated UniScore using a benchmark list of 1106 known essential proteins for *Saccharomyces cerevisiae* and obtained improved predictions as compared to individual CM-based prediction methods. In a more comprehensive study, Rio et al. [30] have used 16 well-known CMs to rank and find essential genes in 18 reconstructed metabolic networks of *Saccharomyces cerevisiae*. They showed that the simultaneous consideration of at least two CMs could lead to a significant identification of essential genes. In this work, we

have used a more recent score (IVI) [297] which has the advantage of including three novel CMs, Collective influence, Local H index, and ClusterRank and combining them in synergistic way with widely used and well-known CMs (Degree, Betweenness and Neighborhood connectivity). The fact that IVI simultaneously takes into consideration global, semi-local and local CMs makes it a suitable metric for capturing all the topological characteristics of the reaction network.

Combination of constraint-based and network-based methods results in a better selection of relevant biochemical reactions that the yeast uses to achieve a given function and metabolic outcome. These reactions and their related genes can be biotechnologically manipulated in order to create modified yeast species that can be used to enhance food aspects, mainly, flavor and safety.

Chapter 5

Conclusion

Biological processes are complex and dynamic. To study them, there is a need for modeling approaches that take into account this complexity. Throughout this thesis, we opted for network-based methods which provide a comprehensive representation of cell complexity, i.e. a representation accounting for the different interactions that can exist between bio-molecules such as genes, proteins and metabolites. The network-based approaches we used, i.e. network diffusion, community detection and centrality measures, allowed us to achieve relevant and promising results over different analysis purposes.

On one hand, the network diffusion applied to *E. coli* protein-protein network has successfully identified new genes related to *E. coli's* antimicrobial resistance to widely used antimicrobial's. Furthermore, some of these genes were also experimentally validated in collaboration with our partners of the E-MUSE consortium. On the other hand, network-based centrality measures, combined with the dynamic constraint-based modeling approaches applied to genome scale metabolic models (GSMM), has permitted the development of a new tool that can be used for the reduction and the construction of context-specific metabolic networks in yeast cells. In terms of their application to cope with real-life challenges, the newly identified genes can be investigated using further *in vitro* assays to develop more effective therapeutic strategies against antimicrobial resistance in *E. coli*. On the other hand, the new GSMM reduction method can provide more relevant yeast metabolic reconstructions that can uncover new metabolic routes under a given condition. These routes can be used to control and redirect the metabolic outcomes of yeasts towards desired phenotypic traits, especially in food industry, such as the enhancement of food flavor profiles and food safety.

Alongside network-based approaches, we also considered applying machine learning methods to analyze cheese-related data collected by our colleagues within the E-MUSE consortium. The work was challenging as the data were sparse, with very few samples and a high number of variables, so not very well suited for machine learning tasks. However, the objective of the work was worth it. We demonstrated a strong microbiome-metabolome relationship in cheese, by using a microbial metatranscriptome dataset for training and another one as independent validation, and made use of it to estimate the cheese flavor profiles directly from the microbial gene expression. Although the challenging data we started from, the performances of the predictive models were, overall, satisfying.

Future Research Perspectives

Flavor as well as texture and, of course, safety, are the cheese's characteristics industrials give more importance to, thus the use of quality control protocols to monitor cheese (off)flavors and contamination is mandatory. We think that assisting experimental quality control methods with *in silico* predictive models can reduce costs and time required to monitor cheese quality and safety. However, training robust predictive models requires large data sets, which is not often available in omics data related to cheese, so, there is a need for a comprehensive literature screening to collect such data. We think that the creation of a "Cheesomics" databases will allow applying

more sophisticated machine learning tools that can be used to improve cheese making processes.

Chapter 6

Dissemination

7th International ISEKI-Food Conference, 5-7 July 2023, Palaiseau, France. Oral presentation of the work: "Estimation of metabolite levels in cheese from microbial gene expression". [311]

Italian Chapter of Complex Systems Society 2023 Conference, 9-11 October 2023, Naples, Italy. Poster presentation of the work: "Network diffusion analysis to elucidate antimicrobial resistance mechanisms of E. coli and reveal potential drug targets". [312]

7TH INTERNATIONAL CONFERENCE ON FOODOMICS, 14-16 February 2024, Cesena, Italy. Poster presentation of the work: "Estimation of metabolite levels in cheese from microbial gene expression". [313]

The 8th Antimicrobial Resistance Conference, 5-9 March 2024, Basel, Switzerland. Poster presentation of the work: "Network diffusion analysis to elucidate antimicrobial resistance mechanisms of E. coli and reveal potential drug targets". [314]

Foundations of Systems Biology in Engineering 2024 Conference, 7-12 September, 2024, Corfu, Greece. Poster presentation of the work: "Network diffusion analysis to elucidate antimicrobial resistance mechanisms of E. coli and reveal potential drug targets". [315]

IDF World Dairy Summit 2024 and EMUSE x FAIROMICS workshop, 14-20 October, 2024, Paris. Oral presentation of the works: "Estimation of metabolite levels in cheese from microbial gene expression", "Network diffusion analysis to elucidate antimicrobial resistance mechanisms of *E. coli* and reveal potential drug targets" and "Context-Specific Genome-Scale Constrained Models Using Transcriptomics, Flux Variability, and Network Topology". [316]

Bibliography

[1] Nerissa Russell. "Milk, wool, and traction: secondary animal products". In: *Ancient Europe* 8000 (2004), pp. 325–333.

[2] Frederick J Simoons. "The antiquity of dairying in Asia and Africa". In: *Geographical Review* (1971), pp. 431–439.

[3] Andrew Dalby. *Cheese: A global history*. Reaktion Books, 2009.

[4] William E Sandine and Paul R Elliker. "Microbially induced flavors and fermented foods. Flavor in fermented dairy products". In: *Journal of Agricultural and Food Chemistry* 18.4 (1970), pp. 557–562.

[5] Juliet Harbutt. *World cheese book*. Dorling Kindersley Ltd, 2015.

[6] Patrick F Fox et al. "Cheese: historical aspects". In: *Fundamentals of cheese science* (2017), pp. 1–10.

[7] Bruna Borges Soares et al. "Chapter 8 - Environmental impact of cheese production". In: *Environmental Impact of Agro-Food Industry and Food Consumption*. Ed. by Charis M. Galanakis. Academic Press, 2021, pp. 169–187. ISBN: 978-0-12-821363-6. DOI: https://doi.org/10.1016/B978-0-12-821363-6.00009-6. URL: https://www.sciencedirect.com/science/article/pii/B9780128213636000096.

[8] Bal Kumari Sharma Khanal, Monica Pradhan, and Nidhi Bansal. "Cheese: Importance and introduction to basic technologies". In: *Journal of Food Science and Technology Nepal* 11 (2019), pp. 14–24.

[9] Patrick F Fox et al. "Overview of cheese manufacture". In: *Fundamentals of cheese science* (2017), pp. 11–25.

[10] Paul LH McSweeney. "Biochemistry of cheese ripening". In: *International journal of dairy technology* 57.2-3 (2004), pp. 127–144.

[11] Paul D Cotter and Tom P Beresford. "Microbiome changes during ripening". In: *Cheese*. Elsevier, 2017, pp. 389–409.

[12] Andrea Biolatto et al. "Seasonal variation in the odour characteristics of whole milk powder". In: *Food Chemistry* 103.3 (2007), pp. 960–967.

[13] Antihus Hernández Gómez et al. "Electronic nose technique potential monitoring mandarin maturity". In: *Sensors and Actuators B: Chemical* 113.1 (2006), pp. 347–353.

[14] Figen Korel and MÖ Balaban. "Microbial and sensory assessment of milk with an electronic nose". In: *Journal of Food Science* 67.2 (2002), pp. 758–764.

[15] Roya Afshari et al. "New insights into cheddar cheese microbiota-metabolome relationships revealed by integrative analysis of multi-omics data". In: *Scientific Reports* 10.1 (2020), p. 3164.

[16] Roya Afshari et al. "Microbiota and metabolite profiling combined with integrative analysis for differentiating cheeses of varying ripening ages". In: *Frontiers in Microbiology* 11 (2020), p. 592060.

[17] Andrea S Bertuzzi et al. "Omics-based insights into flavor development and microbial succession within surface-ripened cheese". In: *MSystems* 3.1 (2018), pp. 10–1128.

[18] Leonhard Euler. “Solutio problematis ad geometriam situs pertinentis”. In: *Commentarii academiae scientiarum Petropolitanae* (1741), pp. 128–140.

[19] Leonhard Euler. “Leonhard Euler and the Königsberg bridges”. In: *Scientific American* 189.1 (1953), pp. 66–72.

[20] Kane O Pryor and Jamie Sleigh. “The seven bridges of Königsberg”. In: *The Journal of the American Society of Anesthesiologists* 114.4 (2011), pp. 739–740.

[21] Dehnokhalaji Akram and Nasrabadi Nasim. “Graph Theory”. In: *Networks of Networks in Biology: Concepts, Tools and Applications*. Ed. by Narsis A. Kiani, David Gomez-Cabrero, and GinestraEditors Bianconi. Cambridge University Press, 2021, 18–32.

[22] Elsevier eBooks. *Bipartite Graph*. 2014. URL: https://www.sciencedirect.com/topics/computer-science/bipartite-graph.

[23] Pietro Hiram Guzzi and Swarup Roy. “2 - Preliminaries of graph theory”. In: *Biological Network Analysis*. Ed. by Pietro Hiram Guzzi and Swarup Roy. Academic Press, 2020, pp. 7–24. ISBN: 978-0-12-819350-1. DOI: https://doi.org/10.1016/B978-0-12-819350-1.00008-6. URL: https://www.sciencedirect.com/science/article/pii/B9780128193501000086.

[24] Harmanjit Singh and Richa Sharma. “Role of adjacency matrix & adjacency list in graph theory”. In: *International Journal of Computers & Technology* 3.1 (2012), pp. 179–183.

[25] Alex Bavelas. “Communication patterns in task-oriented groups”. In: *The journal of the acoustical society of America* 22.6 (1950), pp. 725–730.

[26] Akrati Saxena and Sudarshan Iyengar. “Centrality measures in complex networks: A survey”. In: *arXiv preprint arXiv:2011.07190* (2020).

[27] Xingyi Li et al. “Network-based methods for predicting essential genes or proteins: a survey”. In: *Briefings in bioinformatics* 21.2 (2020), pp. 566–583.

[28] Paolo Boldi and Sebastiano Vigna. “Axioms for centrality”. In: *Internet Mathematics* 10.3-4 (2014), pp. 222–262.

[29] Minoo Ashtiani et al. “A systematic survey of centrality measures for protein-protein interaction networks”. In: *BMC systems biology* 12 (2018), pp. 1–17.

[30] Gabriel del Rio, Dirk Koschützki, and Gerardo Coello. “How to identify essential genes from molecular networks?” In: *BMC systems biology* 3 (2009), pp. 1–12.

[31] John M Bolland. “Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks”. In: *Social networks* 10.3 (1988), pp. 233–253.

[32] Marvin E Shaw. “Group structure and the behavior of individuals in small groups”. In: *The Journal of psychology* 38.1 (1954), pp. 139–149.

[33] LC Freeman. “A set of measures of centrality based on betweenness”. In: *Sociometry* (1977).

[34] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’networks”. In: *nature* 393.6684 (1998), pp. 440–442.

[35] R Duncan Luce and Albert D Perry. “A method of matrix analysis of group structure”. In: *Psychometrika* 14.2 (1949), pp. 95–116.

[36] Duan-Bing Chen et al. “Identifying influential nodes in large-scale directed networks: the role of clustering”. In: *PloS one* 8.10 (2013), e77455.

[37] Jorge E Hirsch. “An index to quantify an individual’s scientific research output”. In: *Proceedings of the National academy of Sciences* 102.46 (2005), pp. 16569–16572.

[38] Jorge E Hirsch. “An index to quantify an individual’s scientific research output that takes into account the effect of multiple coauthorship”. In: *Scientometrics* 85.3 (2010), pp. 741–754.

[39] András Korn, András Schubert, and András Telcs. “Lobby index in networks”. In: *Physica A: Statistical Mechanics and its Applications* 388.11 (2009), pp. 2221–2226.

[40] Linyuan Lü et al. “The H-index of a network node and its relation to degree and coreness”. In: *Nature communications* 7.1 (2016), p. 10168.

[41] Linyuan Lü et al. “Vital nodes identification in complex networks”. In: *Physics reports* 650 (2016), pp. 1–63.

[42] Qiang Liu et al. “Leveraging local h-index to identify and rank influential spreaders in networks”. In: *Physica A: Statistical Mechanics and its Applications* 512 (2018), pp. 379–391. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2018.08.053. URL: https://www.sciencedirect.com/science/article/pii/S0378437118309932.

[43] Sergei Maslov and Kim Sneppen. “Specificity and Stability in Topology of Protein Networks”. In: *Science* 296.5569 (2002), pp. 910–913. DOI: 10.1126/science.1065103. eprint: https://www.science.org/doi/pdf/10.1126/science.1065103. URL: https://www.science.org/doi/abs/10.1126/science.1065103.

[44] Flaviano Morone and Hernán A Makse. “Influence maximization in complex networks through optimal percolation”. In: *Nature* 524.7563 (2015), pp. 65–68.

[45] Linton C. Freeman. “Centrality in social networks conceptual clarification”. In: *Social Networks* 1.3 (1978), pp. 215–239. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(78)90021-7. URL: https://www.sciencedirect.com/science/article/pii/0378873378900217.

[46] Jiaoe Wang et al. “Exploring the network structure and nodal centrality of China’s air transport network: A complex network approach”. In: *Journal of Transport Geography* 19.4 (2011), pp. 712–721.

[47] Arzucan Özgür et al. “Identifying gene-disease associations using centrality on a literature mined gene-interaction network”. In: *Bioinformatics* 24.13 (July 2008), pp. i277–i285. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btn182. eprint: https://academic.oup.com/bioinformatics/article-pdf/24/13/i277/49052827/bioinformatics_24_13_i277.pdf. URL: https://doi.org/10.1093/bioinformatics/btn182.

[48] Adriana Di Martino et al. “Shared and distinct intrinsic functional network centrality in autism and attention-deficit/hyperactivity disorder”. In: *Biological psychiatry* 74.8 (2013), pp. 623–632.

[49] Jörg Menche et al. “Uncovering disease-disease relationships through the incomplete interactome”. In: *Science* 347.6224 (2015), p. 1257601.

[50] Sebastian Köhler et al. “Walking the interactome for prioritization of candidate disease genes”. In: *The American Journal of Human Genetics* 82.4 (2008), pp. 949–958.

[51] Matteo Bersanelli et al. “Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules”. In: *Scientific Reports* 6.1 (2016), p. 34841.

[52] Narsis A Kiani and Mikko Kivelä. “3 Structural Analysis of Biological Networks”. In: *Networks of Networks in Biology: Concepts, Tools and Applications* (2021), p. 35.

[53] Santo Fortunato. “Community detection in graphs”. In: *Physics Reports* 486.3 (2010), pp. 75–174. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2009.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0370157309002841.

[54] Stuart A Rice. “The identification of blocs in small political bodies”. In: *American Political Science Review* 21.3 (1927), pp. 619–627.

[55] Robert S Weiss and Eugene Jacobson. “A method for the analysis of the structure of complex organizations”. In: *American Sociological Review* 20.6 (1955), pp. 661–668.

[56] Mark EJ Newman and Michelle Girvan. “Finding and evaluating community structure in networks”. In: *Physical review E* 69.2 (2004), p. 026113.

[57] Santo Fortunato. “Community detection in graphs”. In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[58] Santo Fortunato and Darko Hric. “Community detection in networks: A user guide”. In: *Physics reports* 659 (2016), pp. 1–44.

[59] Srinivasan Parthasarathy, Yiye Ruan, and Venu Satuluri. “Community discovery in social networks: Applications, methods and emerging trends”. In: *Social network data analytics* (2011), pp. 79–113.

[60] Zhao Yang, René Algesheimer, and Claudio J Tessone. “A comparative analysis of community detection algorithms on artificial networks”. In: *Scientific reports* 6.1 (2016), p. 30750.

[61] Alex Pothen. “Graph Partitioning Algorithms with Applications to Scientific Computing”. In: *Parallel Numerical Algorithms*. Ed. by David E. Keyes, Ahmed Sameh, and V. Venkatakrishnan. Dordrecht: Springer Netherlands, 1997, pp. 323–368. ISBN: 978-94-011-5412-3. DOI: 10.1007/978-94-011-5412-3_12. URL: https://doi.org/10.1007/978-94-011-5412-3_12.

[62] Brian W Kernighan and Shen Lin. “An efficient heuristic procedure for partitioning graphs”. In: *The Bell system technical journal* 49.2 (1970), pp. 291–307.

[63] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[64] Mark EJ Newman. “Fast algorithm for detecting community structure in networks”. In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 69.6 (2004), p. 066133.

[65] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[66] Michelle Girvan and Mark EJ Newman. “Community structure in social and biological networks”. In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[67] N Deborah Friedman et al. “Health care–associated bloodstream infections in adults: a reason to change the accepted definition of community-acquired infections”. In: *Annals of internal medicine* 137.10 (2002), pp. 791–797.

[68] Mohammed Uddin Rasheed et al. “Antimicrobial drug resistance in strains of Escherichia coli isolated from food sources”. In: *Revista do Instituto de Medicina Tropical de São Paulo* 56.4 (2014), pp. 341–346.

[69] Adrian Woolfson. *Origins of life: An improbable journey*. 2015.

[70] Lindsay Morrison and Teresa R Zembower. “Antimicrobial resistance”. In: *Gastrointestinal Endoscopy Clinics* 30.4 (2020), pp. 619–635.

[71] SJ Dancer, P Shears, and DJ Platt. “Isolation and characterization of coliforms from glacial ice and water in Canada’s High Arctic”. In: *Journal of Applied Microbiology* 82.5 (1997), pp. 597–609.

[72] Vanessa M D’Costa et al. “Antibiotic resistance is ancient”. In: *Nature* 477.7365 (2011), pp. 457–461.

[73] URL: https://www.who.int/publications/i/item/no-time-to-wait-securing-the-future-from-drug-resistant-infections.

[74] Douglas W MacPherson et al. “Population mobility, globalization, and antimicrobial drug resistance”. In: *Emerging infectious diseases* 15.11 (2009), p. 1727.

[75] Dongeun Yong et al. “Characterization of a new metallo-β-lactamase gene, bla NDM-1, and a novel erythromycin esterase gene carried on a unique genetic structure in Klebsiella pneumoniae sequence type 14 from India”. In: *Antimicrobial agents and chemotherapy* 53.12 (2009), pp. 5046–5054.

[76] Jack W Scannell et al. “Diagnosing the decline in pharmaceutical R&D efficiency”. In: *Nature reviews Drug discovery* 11.3 (2012), pp. 191–200.

[77] Derya Aytan-Aktug et al. “Predicting antimicrobial resistance using partial genome alignments”. In: *Msystems* 6.3 (2021), pp. 10–1128.

[78] Bilal Shaker et al. “In silico methods and tools for drug discovery”. In: *Computers in Biology and Medicine* 137 (2021), p. 104851. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104851. URL: https://www.sciencedirect.com/science/article/pii/S0010482521006454.

[79] Nagarjuna Prakash Dalbanjan and SK Praveen Kumar. “A Chronicle Review of In-Silico Approaches for Discovering Novel Antimicrobial Agents to Combat Antimicrobial Resistance”. In: *Indian Journal of Microbiology* (2024), pp. 1–15.

[80] Alina Cărunta, Mihai Pleşu, and Mircea Marin. “Antimicrobial Resistance Patterns Detection Using Gene Interaction Networks Analysis”. In: *2019 E-Health and Bioengineering Conference (EHB)*. 2019, pp. 1–4. DOI: 10.1109/EHB47216.2019.8970092.

[81] Aikaterini Sakagianni et al. “Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review”. In: *Antibiotics* 12.3 (2023). ISSN: 2079-6382. DOI: 10.3390/antibiotics12030452. URL: https://www.mdpi.com/2079-6382/12/3/452.

[82] Sravan Kumar Miryala and Sudha Ramaiah. “Exploring the multi-drug resistance in Escherichia coli O157:H7 by gene interaction network: A systems biology approach”. In: *Genomics* 111.4 (2019), pp. 958–965. ISSN: 0888-7543. DOI: https://doi.org/10.1016/j.ygeno.2018.06.002. URL: https://www.sciencedirect.com/science/article/pii/S0888754318302660.

[83] Joel L Sussman et al. “Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules”. In: *Acta Crystallographica Section D: Biological Crystallography* 54.6 (1998), pp. 1078–1084.

[84] Barbara Zdrazil et al. “The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods”. In: *Nucleic acids research* 52.D1 (2024), pp. D1180–D1192.

[85] Sunghwan Kim et al. “PubChem 2023 update”. In: *Nucleic acids research* 51.D1 (2023), pp. D1373–D1380.

[86] Craig Knox et al. “DrugBank 6.0: the DrugBank knowledgebase for 2024”. In: *Nucleic acids research* 52.D1 (2024), pp. D1265–D1275.

[87] Brian K Shoichet. “Virtual screening of chemical libraries”. In: *Nature* 432.7019 (2004), pp. 862–865.

[88] Richa Mishra et al. “Molecular modeling, QSAR analysis and antimicrobial properties of Schiff base derivatives of isatin”. In: *Journal of Molecular Structure* 1243 (2021), p. 130763.

[89] Jonathan M Stokes et al. “A deep learning approach to antibiotic discovery”. In: *Cell* 180.4 (2020), pp. 688–702.

[90] Shota Higashihira et al. “Halicin remains active against Staphylococcus aureus in biofilms grown on orthopaedically relevant substrates”. In: *Bone & Joint Research* 13.3 (2024), p. 101.

[91] August Allen and Lina Nilsson. “The Drug Factory: Industrializing How New Drugs Are Found”. In: *SLAS DISCOVERY: Advancing the Science of Drug Discovery* 26.9 (2021), pp. 1076–1078.

[92] D Aytan-Aktug et al. “Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks”. In: *Msystems* 5.1 (2020), pp. 10–1128.

[93] Katherine E Goodman et al. “A clinical decision tree to predict whether a bacteremic patient is infected with an extended-spectrum β-lactamase–producing organism”. In: *Clinical Infectious Diseases* 63.7 (2016), pp. 896–903.

[94] Sanjat Kanjilal et al. “A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection”. In: *Science translational medicine* 12.568 (2020), eaay5067.

[95] Artur Kadurin et al. “druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico”. In: *Molecular pharmaceutics* 14.9 (2017), pp. 3098–3104.

[96] Yiran Huang and Cheng Zhong. “Detecting list-colored graph motifs in biological networks using branch-and-bound strategy”. In: *Computers in Biology and Medicine* 107 (2019), pp. 1–9. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2019.01.025. URL: https://www.sciencedirect.com/science/article/pii/S0010482519300319.

[97] Ron Milo et al. “Network motifs: simple building blocks of complex networks”. In: *Science* 298.5594 (2002), pp. 824–827.

[98] Vincent Lacroix, Cristina G Fernandes, and Marie-France Sagot. “Motif search in graphs: application to metabolic networks”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 3.4 (2006), pp. 360–368.

[99] Alina Carunta and Mihai Plesu. “Motif Detection in Biological Networks”. In: *2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. 2019, pp. 341–342. DOI: 10.1109/SYNASC49474.2019.00057.

[100] Sravan Kumar Miryala, Anand Anbarasu, and Sudha Ramaiah. “Systems biology studies in Pseudomonas aeruginosa PA01 to understand their role in biofilm formation and multidrug efflux pumps”. In: *Microbial Pathogenesis* 136 (2019), p. 103668. ISSN: 0882-4010. DOI: https://doi.org/10.1016/j.micpath.2019.103668. URL: https://www.sciencedirect.com/science/article/pii/S0882401019304814.

[101] P Anitha, Anand Anbarasu, and Sudha Ramaiah. “Computational gene network study on antibiotic resistance genes of Acinetobacter baumannii”. In: *Computers in Biology and Medicine* 48 (2014), pp. 17–27.

[102] Jeremy S Edwards and Bernhard O Palsson. “Systems properties of the Haemophilus influenzaeRd metabolic genotype”. In: *Journal of Biological Chemistry* 274.25 (1999), pp. 17410–17416.

[103] Ines Thiele and Bernhard Ø Palsson. “A protocol for generating a high-quality genome-scale metabolic reconstruction”. In: *Nature protocols* 5.1 (2010), pp. 93–121.

[104] Adam M Feist et al. “Reconstruction of biochemical networks in microorganisms”. In: *Nature Reviews Microbiology* 7.2 (2009), pp. 129–143.

[105] Christof Francke, Roland J Siezen, and Bas Teusink. “Reconstructing the metabolic network of a bacterium from its genome”. In: *Trends in microbiology* 13.11 (2005), pp. 550–558.

[106] Xin Fang, Colton J Lloyd, and Bernhard O Palsson. “Reconstructing organisms in silico: genome-scale models and their emerging applications”. In: *Nature Reviews Microbiology* 18.12 (2020), pp. 731–743.

[107] Zachary A King et al. “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models”. In: *Nucleic acids research* 44.D1 (2016), pp. D515–D522.

[108] Minoru Kanehisa et al. “From genomics to chemical genomics: new developments in KEGG”. In: *Nucleic acids research* 34.suppl_1 (2006), pp. D354–D357.

[109] Antje Chang et al. “BRENDA, the ELIXIR core data resource in 2021: new developments and updates”. In: *Nucleic acids research* 49.D1 (2021), pp. D498–D508.

[110] Peter D Karp et al. “The BioCyc collection of microbial genomes and metabolic pathways”. In: *Briefings in bioinformatics* 20.4 (2019), pp. 1085–1093.

[111] Samuel MD Seaver et al. “The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes”. In: *Nucleic acids research* 49.D1 (2021), pp. D575–D588.

[112] Filipe Santos, Joost Boele, and Bas Teusink. “A practical guide to genome-scale metabolic models and their analysis”. In: *Methods in enzymology*. Vol. 500. Elsevier, 2011, pp. 509–532.

[113] Adam P Arkin et al. “KBase: the United States department of energy systems biology knowledgebase”. In: *Nature biotechnology* 36.7 (2018), pp. 566–569.

[114] Oscar Dias et al. “Reconstructing genome-scale metabolic models with merlin”. In: *Nucleic acids research* 43.8 (2015), pp. 3899–3910.

[115] Christopher S Henry et al. “High-throughput generation, optimization and analysis of genome-scale metabolic models”. In: *Nature biotechnology* 28.9 (2010), pp. 977–982.

[116] Sebastián N Mendoza et al. “A systematic assessment of current genome-scale metabolic reconstruction tools”. In: *Genome biology* 20 (2019), pp. 1–20.

[117] José P Faria et al. “Methods for automated genome-scale metabolic model reconstruction”. In: *Biochemical Society Transactions* 46.4 (2018), pp. 931–936.

[118] Barbara M. Bakker et al. “Systems biology from micro-organisms to human metabolic diseases: the role of detailed kinetic models”. In: *Biochemical Society Transactions* 38.5 (Sept. 2010), pp. 1294–1301. ISSN: 0300-5127. DOI: 10.1042/BST0381294. eprint: https://portlandpress.com/biochemsoctrans/article-pdf/38/5/1294/547455/bst0381294.pdf. URL: https://doi.org/10.1042/BST0381294.

[119] Marcel HN Hoefnagel et al. “Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis”. In: *Microbiology* 148.4 (2002), pp. 1003–1013.

[120] David Fell and Athel Cornish-Bowden. *Understanding the control of metabolism*. Vol. 2. Portland press London, 1997.

[121] Amit Varma and Bernhard O Palsson. “Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110”. In: *Applied and environmental microbiology* 60.10 (1994), pp. 3724–3731.

[122] Yong Min et al. “Pathway knockout and redundancy in metabolic networks”. In: *Journal of Theoretical Biology* 270.1 (2011), pp. 63–69.

[123] Laura Gómez-Romero, Karina López-Reyes, and Enrique Hernández-Lemus. “The large scale structure of human metabolism reveals resilience via extensive signaling crosstalk”. In: *Frontiers in Physiology* 11 (2020), p. 588012.

[124] Radhakrishnan Mahadevan and Christophe H Schilling. “The effects of alternate optimal solutions in constraint-based genome-scale metabolic models”. In: *Metabolic engineering* 5.4 (2003), pp. 264–276.

[125] Mohammadreza Yasemi and Mario Jolicoeur. “Modelling cell metabolism: a review on constraint-based steady-state and kinetic approaches”. In: *Processes* 9.2 (2021), p. 322.

[126] Radhakrishnan Mahadevan, Jeremy S Edwards, and Francis J Doyle. "Dynamic flux balance analysis of diauxic growth in Escherichia coli". In: *Biophysical journal* 83.3 (2002), pp. 1331–1340.

[127] F Patrick. *Fundamentals of cheese science*. Springer., 2000.

[128] María Castro-Puyana et al. "Reprint of: Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability". In: *TrAC Trends in Analytical Chemistry* 96 (2017), pp. 62–78.

[129] M Meilgaard. "Descriptive analysis techniques". In: *Sensory evaluation techniques* (1999), pp. 161–172.

[130] TK Singh, MA Drake, and KR Cadwallader. "Flavor of Cheddar cheese: A chemical and sensory perspective". In: *Comprehensive reviews in food science and food safety* 2.4 (2003), pp. 166–189.

[131] PJ Bliss et al. "Odour measurement—factors affecting olfactometry panel performance". In: *Water Science and Technology* 34.3-4 (1996), pp. 549–556.

[132] Hely Tuorila and Erminio Monteleone. "Sensory food science in the changing society: Opportunities, needs, and challenges". In: *Trends in Food Science & Technology* 20.2 (2009), pp. 54–62.

[133] Jonathan E Peelle. "Age-related sensory deficits and their consequences". In: (2019).

[134] Allessandra Fraga Da Ré et al. "Tobacco influence on taste and smell: systematic review of the literature". In: *International archives of otorhinolaryngology* 22.01 (2018), pp. 81–87.

[135] Dandan Pu et al. "Characterization of the key aroma compounds in traditional hunan smoke-cured pork leg (Larou, THSL) by aroma extract dilution analysis (AEDA), odor activity value (OAV), and sensory evaluation experiments". In: *Foods* 9.4 (2020), p. 413.

[136] MTSR Gomes, JABP Oliveira, and EM Gaspar. "Use of sensors in cheese manufacture and quality control". In: *2014 IEEE 9th IberoAmerican Congress on Sensors*. IEEE. 2014, pp. 1–4.

[137] Cornelis Weurman. "Isolation and concentration of volatiles in food odor research". In: *Journal of Agricultural and Food Chemistry* 17.2 (1969), pp. 370–384.

[138] Tong Wang et al. "Predicting metabolomic profiles from microbial composition through neural ordinary differential equations". In: *Nature machine intelligence* 5.3 (2023), pp. 284–293.

[139] Himel Mallick et al. "Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences". In: *Nature communications* 10.1 (2019), p. 3136.

[140] James T Morton et al. "Learning representations of microbe–metabolite interactions". In: *Nature methods* 16.12 (2019), pp. 1306–1314.

[141] Vuong Le et al. "Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome". In: *BMC genomics* 21 (2020), pp. 1–15.

[142] Derek Reiman, Brian T Layden, and Yang Dai. "MiMeNet: Exploring microbiome-metabolome relationships using neural networks". In: *PLoS computational biology* 17.5 (2021), e1009021.

[143] Eric Dugat-Bony et al. "Overview of a surface-ripened cheese community functioning by meta-omics analyses". In: *PloS one* 10.4 (2015), e0124360.

[144] James H Bullard et al. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". In: *BMC bioinformatics* 11 (2010), pp. 1–13.

[145] Heiner Klingenberg and Peter Meinicke. “How to normalize metatranscriptomic count data for differential expression analysis”. In: *PeerJ* 5 (2017), e3859.

[146] Mark D Robinson and Alicia Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data”. In: *Genome biology* 11 (2010), pp. 1–9.

[147] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *bioinformatics* 26.1 (2010), pp. 139–140.

[148] Toni Gabaldón and Eugene V Koonin. “Functional and evolutionary implications of gene orthology”. In: *Nature Reviews Genetics* 14.5 (2013), pp. 360–366.

[149] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[150] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (2005), pp. 301–320.

[151] Trevor J Hastie. “Generalized additive models”. In: *Statistical models in S*. Routledge, 2017, pp. 249–307.

[152] Benaglia Tatiana et al. “mixtools: an R package for analyzing finite mixture models”. In: *Journal of Statistical Software* 32.6 (2009), pp. 1–29.

[153] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society: series B (methodological)* 39.1 (1977), pp. 1–22.

[154] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[155] J Macqueen. “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*. 1967.

[156] D Morgenstern and Richard Bellman. “Adaptive control processes: a guided tour”. In: *Econometrica* 30.3 (1962), p. 599.

[157] Gordon Hughes. “On the mean accuracy of statistical pattern recognizers”. In: *IEEE transactions on information theory* 14.1 (1968), pp. 55–63.

[158] Sheng Chen and Haibo He. “Nonstationary stream data learning with imbalanced class distribution”. In: *Imbalanced learning: Foundations, algorithms, and applications* (2013), pp. 151–186.

[159] Donald E Hilt and Donald W Seegrist. *Ridge, a computer program for calculating ridge regression estimates*. Department of Agriculture, Forest Service, Northeastern Forest Experiment . . ., 1977.

[160] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[161] Michel Lutz and Eric Biernat. *Data science: fondamentaux et études de cas: Machine learning avec Python et R*. Editions Eyrolles, 2015.

[162] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[163] T Hastie. “An Introduction to glmnet”. In: *Vignette document of the R glmnet package* (2021).

[164] Max Kuhn. “Building predictive models in R using the caret package”. In: *Journal of statistical software* 28 (2008), pp. 1–26.

[165] Miao Liu et al. “Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar”. In: *Sensors and Actuators B: Chemical* 177 (2013), pp. 970–980.

[166] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[167] Leonard A Breslow, David W Aha, et al. “Simplifying decision trees: A survey”. In: *Knowledge engineering review* 12.1 (1997), pp. 1–40.

[168] Eric Biernat and Michel Lutz. *Data science: fondamentaux et études de cas: Machine learning avec Python et R*. Editions Eyrolles, 2015.

[169] URL: https://statquest.org.

[170] Charles Spearman. “The proof and measurement of association between two things.” In: (1961).

[171] Peter Langfelder and Steve Horvath. “WGCNA: an R package for weighted correlation network analysis”. In: *BMC bioinformatics* 9 (2008), pp. 1–13.

[172] Tianzhi Wu et al. “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data”. In: *The innovation* 2.3 (2021).

[173] Ayesha Judge and Michael S Dodd. “Metabolism”. en. In: *Essays Biochem* 64.4 (Oct. 2020), pp. 607–647.

[174] Şeref Kerem Çorbacıoğlu and Gökhan Aksel. “Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value”. In: *Turkish Journal of Emergency Medicine* 23.4 (2023), p. 195.

[175] Patrick F Fox et al. *Cheese: Chemistry, physics and microbiology, Volume 1: General aspects*. Elsevier, 2004.

[176] Nathalie Japkowicz. “Assessment metrics for imbalanced learning”. In: *Imbalanced learning: Foundations, algorithms, and applications* (2013), pp. 187–206.

[177] César Ferri, José Hernández-Orallo, and R Modroiu. “An experimental comparison of performance measures for classification”. In: *Pattern recognition letters* 30.1 (2009), pp. 27–38.

[178] Arezo Torang, Paraag Gupta, and David J Klinke. “An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets”. In: *BMC bioinformatics* 20 (2019), pp. 1–15.

[179] Murray Moo-Young. *Comprehensive biotechnology*. Elsevier, 2019.

[180] Jens Nielsen. “Metabolic engineering”. In: *Applied microbiology and biotechnology* 55 (2001), pp. 263–283.

[181] Sudheer K. Singh, Syed U. Ahmed, and Ashok Pandey. “Metabolic engineering approaches for lactic acid production”. In: *Process Biochemistry* 41.5 (2006), pp. 991–1000. ISSN: 1359-5113. DOI: https://doi.org/10.1016/j.procbio.2005.12.004. URL: https://www.sciencedirect.com/science/article/pii/S1359511305004782.

[182] Christopher JL Murray et al. “Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis”. In: *The lancet* 399.10325 (2022), pp. 629–655.

[183] URL: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.

[184] Manar Ali Abushaheen et al. “Antimicrobial resistance, mechanisms and its clinical significance”. In: *Disease-a-Month* 66.6 (2020), p. 100971.

[185] Jose M Munita and Cesar A Arias. “Mechanisms of antibiotic resistance”. In: *Virulence mechanisms of bacterial pathogens* (2016), pp. 481–511.

[186] Jean-Pierre Galizzi, Brian Paul Lockhart, and Antoine Bril. "Applying systems biology in drug discovery and development". In: *Drug metabolism and drug interactions* 28.2 (2013), pp. 67–78.

[187] Elaine L Leung et al. "Network-based drug discovery by integrating systems biology and computational technologies". In: *Briefings in bioinformatics* 14.4 (2013), pp. 491–505.

[188] Joaquin Dopazo. "Genomics and transcriptomics in drug discovery". In: *Drug discovery today* 19.2 (2014), pp. 126–132.

[189] Damian Szklarczyk et al. "The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest". In: *Nucleic acids research* 51.D1 (2023), pp. D638–D646.

[190] Jason Montojo et al. "GeneMANIA: Fast gene network construction and function prediction for Cytoscape". In: *F1000Research* 3 (2014).

[191] Christoph Ogris et al. "FunCoup 4: new species, data, and visualization". In: *Nucleic acids research* 46.D1 (2018), pp. D601–D607.

[192] Atanas Kamburov et al. "ConsensusPathDB—a database for integrating human functional interaction networks". In: *Nucleic acids research* 37.suppl_1 (2009), pp. D623–D628.

[193] Noemi Di Nanni et al. "Network diffusion promotes the integrative analysis of multiple omics". In: *Frontiers in genetics* 11 (2020), p. 106.

[194] Stephen Oliver. "Guilt-by-association goes global". In: *Nature* 403.6770 (2000), pp. 601–602.

[195] Lenore Cowen et al. "Network propagation: a universal amplifier of genetic associations". In: *Nature Reviews Genetics* 18.9 (2017), pp. 551–562.

[196] Peggy I Wang et al. "RIDDLE: reflective diffusion and local extension reveal functional associations for unannotated gene sets via proximity in a gene network". In: *Genome biology* 13 (2012), pp. 1–13.

[197] Roded Sharan, Igor Ulitsky, and Ron Shamir. "Network-based prediction of protein function". In: *Molecular systems biology* 3.1 (2007), p. 88.

[198] Mark DM Leiserson et al. "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes". In: *Nature genetics* 47.2 (2015), pp. 106–114.

[199] Pakorn Sagulkoo, Apichat Suratanee, and Kitiporn Plaimas. "Immune-related protein interaction network in severe COVID-19 patients toward the identification of key proteins and drug repurposing". In: *Biomolecules* 12.5 (2022), p. 690.

[200] URL: https://version-11-5.string-db.org/cgi/download?sessionId=b5pBVlyhlDaZ&species_text=Escherichia%2Bcoli%2Bstr.%2BK-12%2Bsubstr.%2BMG1655.

[201] Brian P Alcock et al. "CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database". In: *Nucleic acids research* 51.D1 (2023), pp. D690–D699.

[202] Ea Zankari et al. "PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens". In: *Journal of Antimicrobial Chemotherapy* 72.10 (2017), pp. 2764–2768.

[203] Sergio Picart-Armada et al. "diffuStats: an R package to compute diffusion-based scores on biological networks". In: *Bioinformatics* 34.3 (2018), pp. 533–534.

[204] Maintainer Gabor Csardi. "Package 'igraph'". In: *Last accessed* 3.09 (2013), p. 2013.

[205] URL: https://perso.uclouvain.be/vincent.blondel/research/louvain.html.

[206] Guangchuang Yu et al. “clusterProfiler: an R package for comparing biological themes among gene clusters”. In: *Omics: a journal of integrative biology* 16.5 (2012), pp. 284–287.

[207] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[208] Tomoya Baba et al. “Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection”. In: *Molecular systems biology* 2.1 (2006), pp. 2006–0008.

[209] URL: https://clsi.org/.

[210] URL: https://clsi.org/media/tc4b1paf/m10033_samplepages-1.pdf.

[211] Fred C Tenover. “Mechanisms of antimicrobial resistance in bacteria”. In: *The American journal of medicine* 119.6 (2006), S3–S10.

[212] Manuel F Varela et al. “Bacterial resistance to antimicrobial agents”. In: *Antibiotics* 10.5 (2021), p. 593.

[213] Ada Yonath. “Antibiotics targeting ribosomes: resistance, selectivity, synergism, and cellular regulation”. In: *Annu. Rev. Biochem.* 74.1 (2005), pp. 649–679.

[214] Ditlev E Brodersen et al. “The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30S ribosomal subunit”. In: *Cell* 103.7 (2000), pp. 1143–1154.

[215] Andrew P Carter et al. “Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics”. In: *Nature* 407.6802 (2000), pp. 340–348.

[216] Frank Schlünzen et al. “Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria”. In: *Nature* 413.6858 (2001), pp. 814–821.

[217] José Luis Anaya-López, Joel Edmundo López-Meza, and Alejandra Ochoa-Zarzosa. “Bacterial resistance to cationic antimicrobial peptides”. In: *Critical reviews in microbiology* 39.2 (2013), pp. 180–195.

[218] Michael Zasloff. “Antimicrobial peptides of multicellular organisms”. In: *nature* 415.6870 (2002), pp. 389–395.

[219] Hilde Ulvatne et al. “Proteases in Escherichia coli and Staphylococcus aureus confer reduced susceptibility to lactoferricin B”. In: *Journal of Antimicrobial Chemotherapy* 50.4 (2002), pp. 461–467.

[220] Enrique Llobet, Juan M Tomas, and Jose A Bengoechea. “Capsule polysaccharide is a bacterial decoy for antimicrobial peptides”. In: *Microbiology* 154.12 (2008), pp. 3877–3886.

[221] Edward P Abraham and Ernst Chain. “An enzyme from bacteria able to destroy penicillin”. In: *Nature* 146.3713 (1940), pp. 837–837.

[222] N Pandey and M Cascella. “Beta Lactam Antibiotics.[Updated 2022 Feb 5]”. In: *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing* (2022).

[223] John W Campbell and John E Cronan Jr. “Bacterial fatty acid biosynthesis: targets for antibacterial drug discovery”. In: *Annual Reviews in Microbiology* 55.1 (2001), pp. 305–332.

[224] Richard J Heath et al. “Mechanism of triclosan inhibition of bacterial fatty acid synthesis”. In: *Journal of Biological Chemistry* 274.16 (1999), pp. 11110–11114.

[225] Yong-Mei Zhang and Charles O Rock. “Membrane lipid homeostasis in bacteria”. In: *Nature Reviews Microbiology* 6.3 (2008), pp. 222–233.

[226] Arnold S Bayer et al. “In vitro resistance of Staphylococcus aureus to thrombin-induced platelet microbicidal protein is associated with alterations in cytoplasmic membrane fluidity”. In: *Infection and immunity* 68.6 (2000), pp. 3548–3553.

[227] Manuel Pazos and Katharina Peters. “Peptidoglycan”. In: *Bacterial cell walls and membranes* (2019), pp. 127–168.

[228] Jean-Emmanuel Hugonnet et al. “Factors essential for L, D-transpeptidase-mediated peptidoglycan cross-linking and β-lactam resistance in Escherichia coli”. In: *Elife* 5 (2016), e19469.

[229] Vincent Cattoir et al. “Novel chromosomal mutations responsible for fosfomycin resistance in Escherichia coli”. In: *Frontiers in Microbiology* 11 (2020), p. 575031.

[230] Kathryn Beabout et al. “The ribosomal S10 protein is a general target for decreased tigecycline susceptibility”. In: *Antimicrobial agents and chemotherapy* 59.9 (2015), pp. 5561–5566.

[231] Edward L Bolt et al. “Identification of Escherichia coli ygaQ and rpmG as novel mitomycin C resistance factors implicated in DNA repair”. In: *Bioscience Reports* 36.1 (2016), e00290.

[232] Esmeralda Z Reyes-Fernández and Shimon Schuldiner. “Acidification of cytoplasm in Escherichia coli provides a strategy to cope with stress and facilitates development of antibiotic resistance”. In: *Scientific reports* 10.1 (2020), p. 9954.

[233] Kenneth L Roland, Charles R Esther, and John K Spitznagel. “Isolation and characterization of a gene, pmrD, from Salmonella typhimurium that confers resistance to polymyxin when expressed in multiple copies”. In: *Journal of bacteriology* 176.12 (1994), pp. 3589–3597.

[234] Erica J Rubin et al. “PmrD is required for modifications to Escherichia coli endotoxin that promote antimicrobial resistance”. In: *Antimicrobial agents and chemotherapy* 59.4 (2015), pp. 2051–2061.

[235] Miguel Viveiros et al. “Antibiotic stress, genetic response and altered permeability of E. coli”. In: *PloS one* 2.4 (2007), e365.

[236] Myrielle Dupont et al. “Enterobacter aerogenes OmpX, a cation-selective channel mar-and osmo-regulated”. In: *FEBS letters* 569.1-3 (2004), pp. 27–30.

[237] Nicolas Bouquin et al. “Resistance to trifluoroperazine, a calmodulin inhibitor, maps to the fabD locus in Escherichia coli”. In: *Molecular and General Genetics MGG* 246 (1995), pp. 628–637.

[238] Suzanne Jackowski et al. “A missense mutation in the fabB (β-ketoacyl-acyl carrier protein synthase I) gene confers thiolactomycin resistance to Escherichia coli”. In: *Antimicrobial agents and chemotherapy* 46.5 (2002), pp. 1246–1252.

[239] Daina Zeng et al. “Mutants resistant to LpxC inhibitors by rebalancing cellular homeostasis”. In: *Journal of Biological Chemistry* 288.8 (2013), pp. 5475–5486.

[240] Vaishali Humnabadkar et al. “UDP-N-acetylmuramic acid l-alanine ligase (MurC) inhibition in a tolC mutant Escherichia coli strain leads to cell death”. In: *Antimicrobial Agents and Chemotherapy* 58.10 (2014), pp. 6165–6171.

[241] Srijan Ranjitkar et al. “Identification of mutations in the mrdA gene encoding PBP2 that reduce carbapenem and diazabicyclooctane susceptibility of Escherichia coli clinical isolates with mutations in ftsI (PBP3) and which carry bla NDM-1”. In: *Msphere* 4.4 (2019), pp. 10–1128.

[242] Seth P Cohen, H Hächler, and SB193236 Levy. “Genetic and functional analysis of the multiple antibiotic resistance (mar) locus in Escherichia coli”. In: *Journal of bacteriology* 175.5 (1993), pp. 1484–1492.

[243] Emma R Holden et al. “Genome-wide analysis of genes involved in efflux function and regulation within Escherichia coli and Salmonella enterica serovar Typhimurium”. In: *Microbiology* 169.2 (2023), p. 001296.

[244] SW Cowan et al. “Crystal structures explain functional properties of two E. coli porins”. In: *Nature* 358.6389 (1992), pp. 727–733.

[245] Kelly Magnuson et al. “Regulation of fatty acid biosynthesis in Escherichia coli”. In: *Microbiological reviews* 57.3 (1993), pp. 522–542.

[246] Richard J Heath and Charles O Rock. “Roles of the FabA and FabZ β-hydroxyacyl-acyl carrier protein dehydratases in Escherichia coli fatty acid biosynthesis”. In: *Journal of Biological Chemistry* 271.44 (1996), pp. 27795–27801.

[247] Dominique Liger et al. “Over-production, Purification and Properties of the Uridine-diphosphate-N-Acetylmuramate: l-alanine Ligase from Escherichia coli”. In: *European journal of biochemistry* 230.1 (1995), pp. 80–87.

[248] Jesse S Wright III and Robert J Kadner. “The phosphoryl transfer domain of UhpB interacts with the response regulator UhpA”. In: *Journal of bacteriology* 183.10 (2001), pp. 3149–3159.

[249] Melanie A Adams and Zongchao Jia. “Modulator of drug activity B from Escherichia coli: crystal structure of a prokaryotic homologue of DT-diaphorase”. In: *Journal of molecular biology* 359.2 (2006), pp. 455–465.

[250] Melanie A Adams, Pietro Iannuzzi, and Zongchao Jia. “MdaB from Escherichia coli: cloning, purification, crystallization and preliminary X-ray analysis”. In: *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 61.2 (2005), pp. 235–238.

[251] Azar Shahpiri et al. “Enhancement of chromate bioaccumulation by engineered Escherichia coli cells co-expressing chromate reductase (YieF) and a rice metallothionein isoform (OsMT1)”. In: *Journal of Chemical Technology & Biotechnology* 96.5 (2021), pp. 1285–1291.

[252] JD Keasling. “Regulation of intracellular toxic metals and other cations by hydrolysis of polyphosphate.” In: *Annals of the New York Academy of Sciences* 829 (1997), pp. 242–249.

[253] Riko Shirakawa et al. “Knockout of ribosomal protein RpmJ leads to zinc resistance in Escherichia coli”. In: *PloS one* 18.3 (2023), e0277162.

[254] Uwem Okon Edet, Ini Ubi Bassey, and Akaninyene Paul Joseph. “Heavy metal co-resistance with antibiotics amongst bacteria isolates from an open dumpsite soil”. In: *Heliyon* 9.2 (2023).

[255] Elena Anedda et al. “Evaluating the impact of heavy metals on antimicrobial resistance in the primary food production environment: A scoping review”. In: *Environmental Pollution* 320 (2023), p. 121035.

[256] Ting-Ting Qin et al. “SOS response and its regulation on the fluoroquinolone resistance”. In: *Annals of translational medicine* 3.22 (2015).

[257] Arianne M Babina et al. “An S6: S18 complex inhibits translation of E. coli rpsF”. In: *RNA* 21.12 (2015), pp. 2039–2046.

[258] Sheri K Wilcox, Gregory S Cavey, and James D Pearson. “Single ribosomal protein mutations in antibiotic-resistant bacteria analyzed by mass spectrometry”. In: *Antimicrobial agents and chemotherapy* 45.11 (2001), pp. 3046–3055.

[259] Weijie You et al. “Crystallization and preliminary X-ray diffraction analysis of the putative aldose 1-epimerase YeaD from Escherichia coli”. In: *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* 66.8 (2010), pp. 951–953.

[260] STM Allard, M-F Giraud, and James Henderson Naismith. “Epimerases: structure, function and mechanism”. In: *Cellular and Molecular Life Sciences CMLS* 58 (2001), pp. 1650–1665.

[261] Koichi Miura et al. “Molecular cloning of the nemA gene encoding N-ethylmaleimide reductase from Escherichia coli”. In: *Biological and Pharmaceutical Bulletin* 20.1 (1997), pp. 110–112.

[262] Aiqin Shi et al. “Plasmidic expression of nemA and yafC* increased resistance of ethanologenic Escherichia coli LY180 to nonvolatile side products from dilute acid treatment of sugarcane bagasse and artificial hydrolysate”. In: *Applied and Environmental Microbiology* 82.7 (2016), pp. 2137–2145.

[263] Salma Waheed Sheikh et al. “Insights into emergence of antibiotic resistance in acid-adapted enterohaemorrhagic Escherichia coli”. In: *Antibiotics* 10.5 (2021), p. 522.

[264] Alexia N Torres et al. “Deciphering additional roles for the EF-Tu, l-Asparaginase II and OmpT proteins of Shiga toxin-producing Escherichia coli”. In: *Microorganisms* 8.8 (2020), p. 1184.

[265] Alina Cărunta, Mihai Pleşu, and Mircea Marin. “Antimicrobial Resistance Patterns Detection Using Gene Interaction Networks Analysis”. In: *2019 E-Health and Bioengineering Conference (EHB)*. IEEE. 2019, pp. 1–4.

[266] P Anitha, Anand Anbarasu, and Sudha Ramaiah. “Gene network analysis reveals the association of important functional partners involved in antibiotic resistance: a report on an important pathogenic bacterium Staphylococcus aureus”. In: *Gene* 575.2 (2016), pp. 253–263.

[267] M Anusha et al. “Gene network interaction analysis to elucidate the antimicrobial resistance mechanisms in the Clostridium difficile”. In: *Microbial Pathogenesis* 178 (2023), p. 106083.

[268] Pavan Gollapalli, H Manjunatha, Praveenkumar Shetty, et al. “Network topology analysis of essential genes interactome of Helicobacter pylori to explore novel therapeutic targets”. In: *Microbial Pathogenesis* 158 (2021), p. 105059.

[269] Prasanna Kumar Selvam et al. “Decoding the Complex Genetic Network of Antimicrobial Resistance in Campylobacter jejuni using Advanced Gene Network Analysis”. In: (2023).

[270] Thandavarayan Ramamurthy et al. “Deciphering the genetic network and programmed regulation of antimicrobial resistance in bacterial pathogens”. In: *Frontiers in Cellular and Infection Microbiology* 12 (2022), p. 952491.

[271] Hsuan-Lin Her, Po-Ting Lin, and Yu-Wei Wu. “PangenomeNet: a pan-genome-based network reveals functional modules on antimicrobial resistome for Escherichia coli strains”. In: *BMC bioinformatics* 22 (2021), pp. 1–19.

[272] Mengjun Hu et al. “Characterization of the role of two-component systems in antibiotic resistance formation in Salmonella enterica serovar Enteritidis”. In: *Msphere* 7.6 (2022), e00383–22.

[273] Soojin Jang. “AcrAB- TolC, a major efflux pump in Gram negative bacteria: toward understanding its operation mechanism”. In: *BMB reports* 56.6 (2023), p. 326.

[274] Anne Liu et al. “Antibiotic sensitivity profiles determined with an Escherichia coli gene knockout collection: generating an antibiotic bar code”. In: *Antimicrobial agents and chemotherapy* 54.4 (2010), pp. 1393–1403.

[275] MICHAEL D Island, BY Wei, and ROBERT J Kadner. “Structure and function of the uhp genes for the sugar phosphate transport system in Escherichia coli and Salmonella typhimurium”. In: *Journal of bacteriology* 174.9 (1992), pp. 2754–2762.

[276] PC Maloney et al. “Anion-exchange mechanisms in bacteria”. In: *Microbiological reviews* 54.1 (1990), pp. 1–17.

[277] Alap R Subramanian and Eric R Dabbs. “Functional studies on ribosomes lacking protein L1 from mutant Escherichia coli”. In: *European Journal of Biochemistry* 112.2 (1980), pp. 425–430.

[278] Jyoti Tanwar et al. “Multidrug resistance: an emerging crisis”. In: *Interdisciplinary perspectives on infectious diseases* 2014.1 (2014), p. 541340.

[279] Markus W. Covert et al. “Metabolic modeling of microbial strains in silico”. In: *Trends in Biochemical Sciences* 26.3 (2001), pp. 179–186. ISSN: 0968-0004. DOI: https://doi.org/10.1016/S0968-0004(00)01754-0. URL: https://www.sciencedirect.com/science/article/pii/S0968000400017540.

[280] Matthew A Oberhardt, Bernhard Ø Palsson, and Jason A Papin. "Applications of genome-scale metabolic reconstructions". In: *Molecular systems biology* 5.1 (2009), p. 320.

[281] Jeroen Hugenholtz. "The lactic acid bacterium as a cell factory for food ingredient production". In: *International Dairy Journal* 18.5 (2008). Netherlands Association for the Advancement of Dairy Science 1908-2008, pp. 466–475. ISSN: 0958-6946. DOI: https://doi.org/10.1016/j.idairyj.2007.11.015. URL: https://www.sciencedirect.com/science/article/pii/S0958694607002385.

[282] Lucia Brown et al. "Lactic acid bacteria as cell factories for the generation of bioactive peptides". In: *Protein and peptide letters* 24.2 (2017), pp. 146–155.

[283] Martin H. Rau and Ahmad A. Zeidan. "Constraint-based modeling in microbial food biotechnology". In: *Biochemical Society Transactions* 46.2 (Mar. 2018), pp. 249–260. ISSN: 0300-5127. DOI: 10.1042/BST20170268. eprint: https://portlandpress.com/biochemsoctrans/article-pdf/46/2/249/434211/bst-2017-0268c.pdf. URL: https://doi.org/10.1042/BST20170268.

[284] Dennis Schlossarek et al. "Rewiring of the protein–protein–metabolite interactome during the diauxic shift in yeast". In: *Cellular and Molecular Life Sciences* 79.11 (2022), p. 550.

[285] Scott A Becker and Bernhard O Palsson. "Context-specific metabolic networks are consistent with experiments". In: *PLoS computational biology* 4.5 (2008), e1000082.

[286] Maria Pires Pacheco and Thomas Sauter. "The FASTCORE Family: For the Fast Reconstruction of Compact Context-Specific Metabolic Networks Models". In: *Metabolic Network Reconstruction and Modeling: Methods and Protocols*. Ed. by Marco Fondi. New York, NY: Springer New York, 2018, pp. 101–110. ISBN: 978-1-4939-7528-0. DOI: 10.1007/978-1-4939-7528-0_4. URL: https://doi.org/10.1007/978-1-4939-7528-0_4.

[287] Meric Ataman et al. "redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models". In: *PLoS computational biology* 13.7 (2017), e1005444.

[288] Sudhakar Jonnalagadda, Balaji Balagurunathan, and Rajagopalan Srinivasan. "Graph theory augmented math programming approach to identify minimal reaction sets in metabolic networks". In: *Computers Chemical Engineering* 35.11 (2011), pp. 2366–2377. ISSN: 0098-1354. DOI: https://doi.org/10.1016/j.compchemeng.2011.05.006. URL: https://www.sciencedirect.com/science/article/pii/S0098135411001670.

[289] Romain Minebois, Roberto Pérez-Torrado, and Amparo Querol. "A time course metabolism comparison among *Saccharomyces cerevisiae*, *S. uvarum* and *S. kudriavzevii* species in wine fermentation". In: *Food Microbiol.* 90 (2020), p. 103484. DOI: 10.1016/j.fm.2020.103484.

[290] V. Rojas et al. "Studies on acetate ester production by non-*Saccharomyces* wine yeasts". In: *Int J Food Microbiol* 70.3 (2001), pp. 283–289.

[291] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9 (2014), pp. 357–359.

[292] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". In: *Bioinformatics* 31.2 (2014), pp. 166–169.

[293] URL: https://sbml.org/documents/what-is-sbml/.

[294] Ahmed Mohamed et al. "NetPathMiner: R/Bioconductor package for network path mining through gene expression". In: *Bioinformatics* 30.21 (2014), pp. 3139–3141.

[295] Gabor Csardi and Tamas Nepusz. “The igraph software”. In: *Complex syst* 1695 (2006), pp. 1–9.

[296] Per Hage and Frank Harary. “Eccentricity and centrality in networks”. In: *Social Networks* 17.1 (1995), pp. 57–63. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(94)00248-9. URL: https://www.sciencedirect.com/science/article/pii/0378873394002489.

[297] Adrian Salavaty, Mirana Ramialison, and Peter D Currie. “Integrated value of influence: an integrative method for the identification of the most influential nodes within networks”. In: *Patterns* 1.5 (2020).

[298] William Ogilvy Kermack and Anderson G McKendrick. “A contribution to the mathematical theory of epidemics”. In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115.772 (1927), pp. 700–721.

[299] Norman TJ Bailey. *The mathematical theory of infectious diseases and its applications*. 2nd ediition. 1975.

[300] Saumyadipta Pyne, Anile Kumar S. Vullikanti, and Madhav V. Marathe. “Chapter 8 - Big Data Applications in Health Sciences and Epidemiology”. In: *Big Data Analytics*. Ed. by Venu Govindaraju, Vijay V. Raghavan, and C.R. Rao. Vol. 33. Handbook of Statistics. Elsevier, 2015, pp. 171–202. DOI: https://doi.org/10.1016/B978-0-444-63492-4.00008-3. URL: https://www.sciencedirect.com/science/article/pii/B9780444634924000083.

[301] E. Balsa-Canto et al. “AMIGO2, a toolbox for dynamic modeling, optimization and control in systems biology.” In: *Bioinformatics* 32(21) (2016), pp. 3357–3359.

[302] J Schellenberger et al. “Quantitative prediction of cellular metabolism with constraint-based models: the egea Toolbox v2.0.” In: *Nat Protoc. 2011 Aug 4;6(9):1290-307. doi: 10.1038/nprot.2011.308. PMID: 21886097; PMCID: PMC3319681.* (2011).

[303] L. et al. Heirendt. “Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0.” In: *Nat. Protoc.* 14 (2019), 639–702.

[304] Hema Sekhar Reddy Rajula, Matteo Mauri, and Vassilios Fanos. “Scale-free networks in metabolomics”. In: *Bioinformation* 14.3 (2018), p. 140.

[305] URL: https://www.sncf-connect.com/statics/pdf/TRSP-INTER/SVI/Com%20Externe%20Maurienne%20circu%2026-08_FR.pdf.

[306] R Mahadevan and BO Palsson. “Properties of metabolic networks: structure versus function”. In: *Biophysical journal* 88.1 (2005), pp. L07–L09.

[307] Areejit Samal et al. “Low degree metabolites explain essential reactions and enhance modularity in biological networks”. In: *BMC bioinformatics* 7 (2006), pp. 1–10.

[308] Sergey Brin and Lawrence Page. “The anatomy of a large-scale hypertextual web search engine”. In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.

[309] Mariano Beguerisse-Díaz et al. “Flux-dependent graphs for metabolic networks”. In: *NPJ systems biology and applications* 4.1 (2018), p. 32.

[310] Hon Nian Chua et al. “A unified scoring scheme for detecting essential proteins in protein interaction networks”. In: *2008 20th IEEE International Conference on Tools with Artificial Intelligence*. Vol. 2. IEEE. 2008, pp. 66–73.

[311] URL: https://www.dropbox.com/scl/fo/gw06djgvs77n8qtsoxp4q/AG2PUnFfmdp1-LiwceJCFVA/6%20July%20afternoon%20-%20Session%202%20-%20Research?dl=0&preview=6-07+Session+2-9+-+14h55+-+Anis+Mansouri+-unibo-it++-+iseki+2023.pdf&rlkey=tiic1ns6cjq1vqjerdkjvqhvf&subfolder_nav_tracking=1.

[312] URL: https://italy.cssociety.org/index.php/2023/05/23/ccs-italy-conference-2023/.

[313] URL: https://www.foodomics.org.

[314] URL: https://amr-conference.com.

[315] URL: https://fosbe2024.iceht.forth.gr.

[316] URL: https://fil-idf.org/idf_events/world-dairy-summit-2024/.